From Department of Medical Biochemistry and Biophysics
Karolinska Institutet, Stockholm, Sweden

DECODING THE CONTRIBUTION OF TRANSCRIPTION FACTORS TO CELL FATE DETERMINATION

Bei Wei

Karolinska Institutet

Stockholm 2018

Published by Karolinska Institutet.
Printed by Eprint AB 2018

ISBN 978-91-7831-269-6

Decoding the contribution of transcription factors to cell fate determination

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Bei Wei

Principal Supervisor:
Dr. Minna Taipale
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Functional genomics

Co-supervisor:
Professor Jussi Taipale
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Functional genomics

Opponent:
Professor Alexander Stark
Research Institute of Molecular Pathology, Vienna, Austria

Examination Board:
Professor Rickard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology

Professor Johan Elf
Uppsala University
Department of Cell and Molecular Biology

Professor Claes Wadelius
Uppsala University
Department of Immunology, Genetics and Pathology
Division of Medical Genetics and Genomics

Whereof one cannot speak, thereof one must be silent.
- Ludwig Wittgenstein

To my dearest family

致我最爱的家人

ABSTRACT

It is well accepted that transcription factors (TFs) play a crucial role in determining cell identity. Although RNA expression or protein abundance data show that a large fraction of the total TFs are expressed in a given cell, however, only a small set of them is essential for specifying cell identity. This was elegantly demonstrated through reprogramming of somatic cells to induced pluripotent stem cells (iPSCs) by means of ectopic expression of only four key TFs, Oct-3/4 (Pou5f1), Sox2, Klf4 and c-Myc.

In order to decipher the most predominant TFs in specific cell types, we developed a novel massively parallel protein activity assay, Active TF Identification (ATI) that measured DNA-binding activity of TFs in the cell nucleus. This method indicated that around 15 TFs have the highest DNA-binding activities, among which there are "common" TFs universally active in most cell types, "shared" TFs which are active in several cell types and "specific" TFs which are active in only one or two cell types.

It has been well established that the gene transcription is highly correlated with disruption of nucleosomes at the gene regulatory elements. In order to test if TFs are the major determinant of chromatin accessibility, we compared the ATI data with the DNase I hypersensitive sites (DHSs) from the same cell or tissue type, and found out that the enriched subsequences in the ATI results are also enriched within the DHSs compared with the non-DHS regions. This suggested that the DNA-binding activity of TFs, especially the most active ones, played major roles in determining the chromatin accessibility.

In addition, we also performed the ATI assay using nucleosomal DNA to determine the "pioneer" TFs in cells that are capable of binding condensed chromatin.

This study has generated a deeper understanding of the gene regulatory logic and helped us to decipher important TFs in specific cell types.

LIST OF PUBLICATIONS

I. **Wei B**, Jolma A, Sahu B, Orre L M, Zhong F, Zhu F, Kivioja T, Sur I, Lehtiö J, Taipale M and Taipale J. A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. *Nature Biotechnology*. 2018 May 15; 36: 521-529.

II. Zhu F, Farnung L, Kaasinen E, Sahu B, Yin Y, **Wei B**, Dodonova S, Nitta K, Morgunova E, Taipale M, Cramer P and Taipale J. The interaction landscape between transcription factors and the nucleosome. *Nature*. 2018 Sep 24.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
A	adenine
ACN	Acetonitrile
AGC	automatic gain control
AP2	activating protein 2
ATAC-seq	assay for transposase-accessible chromatin using sequencing
ATI	active transcription factor identification
ATP	adenosine triphosphate
bHLH	basic helix-loop-helix
BMP	bone morphogenetic protein
bp	base pair
bZIP	basic leucine zipper
C	cytosine
CBP	CREB-binding protein
cDNA	complementary DNA
CEBP	CCAAT-enhancer-binding proteins
CHD	chromodomain-helicase DNA-binding
ChIP	chromatin immunoprecipitation
CREB	cAMP response element-binding protein
DBD	DNA binding domain
DHS	DNase I hypersensitive site
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNase	Deoxyribonuclease
DNMT	DNA methyltransferase
dsDNA	double stranded DNA
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EGTA	ethylene glycol tetraacetic acid
EMSA	electrophoretic mobility shift assay
ENCODE	encyclopedia of DNA elements
ETS	E26 transformation-specific
FA	formic acid

FAIRE	formaldehyde-assisted isolation of regulatory elements
FDR	false discovery rate
FTMS	fourier transform mass spectrometry
G	Guanine
GTF	general transcription factor
H3K36me3	Histone 3 lysine 36 trimethylation
H3K4me3	Histone 3 lysine 4 trimethylation
H3K9ac	Histone 3 lysine 9 acetylation
H3K9me	Histone 3 lysine 9 methylation
HCD	higher-energy collision dissociation
HCM	hepatocyte culture medium
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HT-SELEX	high-throughput systematic evolution of ligands by exponential enrichment
IAA	iodoacetic acid
IEF	isoelectric focusing
iHep	induced hepatocyte
INO80	inositol-requiring 80
IPG	immobilized pH gradient
iPSC	induced pluripotent stem cell
ISWI	imitation switch
KAT	lysine acetyltransferase
kb	kilo base pair
K_d	equilibrium dissociation constant
KDAC	lysine deacetylase
LIF	leukemia inhibitory factor
MNase	micrococcal nuclease
MS	mass spectrometry
NCP	nucleosome core particle
NR	nuclear receptor
O-GlcNAc	O-linked beta-N-acetylglucosamine
ORF	open reading frame
PBM	protein binding microarray
PCR	polymerase chain reaction
PIC	pre-initiation complex
poly-dA	poly-deoxyadenylic acid
poly-dIdC	poly-deoxy-inosinic-deoxy-cytidylic acid

PRC2	polycomb repressive complex 2
PSM	peptide spectrum match
PWM	position weight matrix
QE	Q-exactive
RBPJ	recombining binding protein suppressor of hairless
RNA	ribonucleic acid
RSC	chromatin structure remodeling
RT-PCR	Real-time PCR
SBP	streptavidin binding peptide
SELEX	systematic evolution of ligands by exponential enrichment
SNP	single nucleotide polymorphism
ssDNA	single stranded DNA
SWI/SNF	switch/sucrose non-fermentable
T	Thymine
T-ALL	T-cell acute lymphoblastic leukemia
TEG	triethylene glycol spacer
TET	ten-eleven translocation
TF	transcription factor
VEGF	vascular endothelial growth factor
WWW	World Wide Web
Y1H	yeast one-hybrid
Y2H	yeast two-hybrid

1 INTRODUCTION

1.1 TRANSCRIPTION FACTORS

Transcription factors are proteins that can bind specific DNA sequences through their DNA-binding domains. It is well established that transcription factors (TFs) play crucial roles in determining specific cell identity[1-4], and that a large fraction of all TFs are actually expressed in most cell types[5,6]. However, it is still unclear how TFs interact with each other as well as other regulators to set up the whole transcriptome profile in a given lineage.

1.1.1 Basic features of TFs

In general, there are two types of TFs, the general transcription factors (GTFs) and the specific TFs. The GTFs are a group of proteins responsible for recognizing the specific sites within the core promoters of transcribed genes and recruiting RNA polymerase II to form the pre-initiation complex (PIC) for the initiation of transcription. In bacteria, there is only one GTF, the sigma factor[7]; in eukaryotes, the GTFs include six members, namely TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, which work together to recruit RNA polymerase II to initiate transcription[8]. Other TFs, known as specific TFs, are capable of interacting with DNA through their DNA binding domains (DBDs) with high specificity. The specific TFs usually bind specific regulatory regions (including promoters and enhancers) of the genome to induce (or repress) transcription of particular genes in different cell types. Compared with the GTFs, the specific TFs account for a vast majority of the transcription factors and possess more diverse DNA-binding specificities. Furthermore, since GTFs are universally expressed in all cell types whereas specific TFs are more restricted in their expression, it is therefore plausible that specific TFs play more important roles in determining the cell identity than the GTFs do. Hereafter, the TFs refer to specific TFs in this thesis.

The transcription factors interact non-covalently with DNA, mainly through hydrogen bonds and Van der Waals forces. The contact between TFs and DNA contains both specific and nonspecific interactions. The specific interactions occur between the specific TF amino acid residues and the nucleotide bases within the core binding sites of the DNA sequences; the nonspecific interactions usually occur between the TF and the phosphate backbone of the DNA[9]. Although the nonspecific interactions do not contribute to the binding specificity of TFs, they are actually quantitatively significant for TF/ DNA binding[10,11].

Binding specificities of more than 1000 TFs from over 131 species have been determined and classified into 54 groups based on their DBDs (**Figure 1.1**)[12], indicating that many TFs share DNA binding specificities with similarly structured DBDs. In addition, the binding specificities of TFs are conserved among distant species; for

instance, almost all *Drosophila* TFs share similar binding motifs with their mammalian orthologous genes[13]. Another important feature of a TF is its binding affinity with DNA. The binding affinity indicates how strong the interaction is between the TF and the corresponding DNA sequence, and it varies among different DBDs, and even among different members within the same TF family. An estimate of the binding affinity is important to understand how TFs target different genomic loci to regulate gene expression. As the binding affinity of specific TF to DNA determines the equilibrium dissociation constant (K_d) of the interaction between the TF and the DNA molecule, it therefore determines the minimum concentration of a TF needed for targeting the specific DNA sequence in the genome. Currently we still lack the knowledge about binding affinity between different kinds of TFs and DNA sequences.

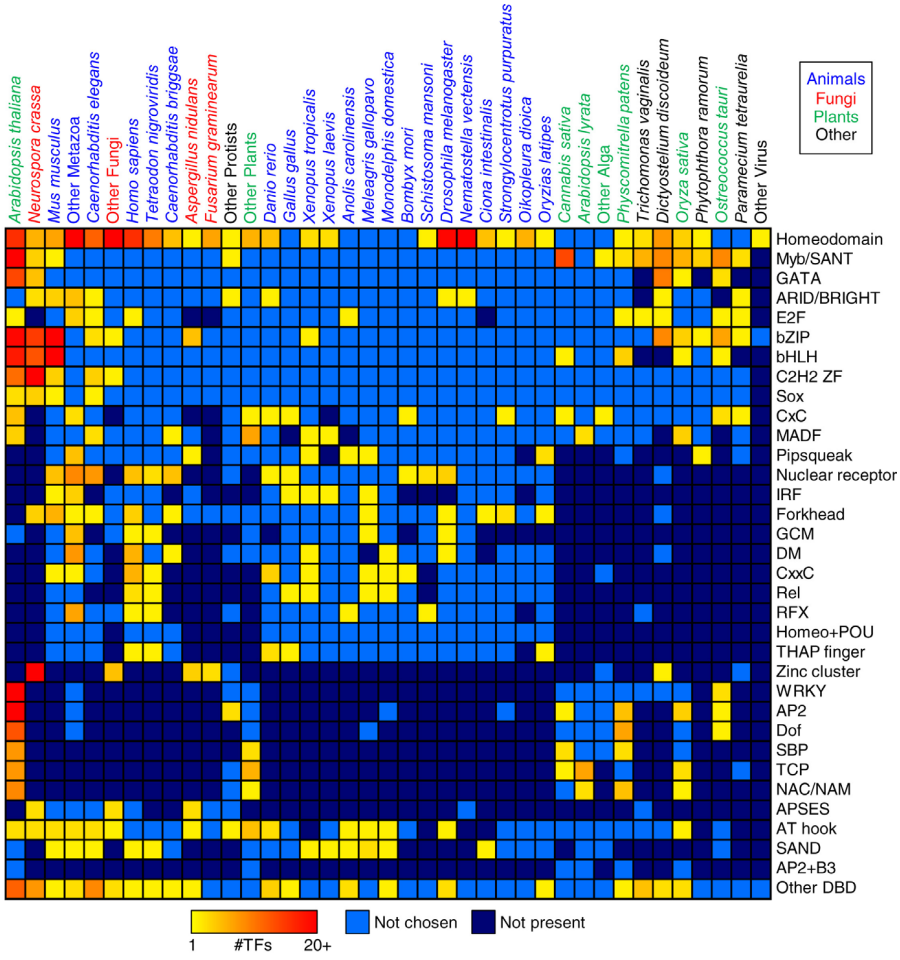

Figure 1.1 Classification of TFs into 54 DBD groups

All of the studied 1,032 TFs from more than 131 species are classified into 54 groups in total based on the DBD classes. The DBD classes or species containing fewer than five members are grouped into "Other". The species are ordered by the total number of TFs with characterized motifs. Note that some TFs have more than one DBD classes, they are therefore classified into independent groups, namely the "Homeo+POU" and the "AP2+B3" groups. Figure is adapted from [12].

Apart from interacting with the DNA sequences on their own, a TF can also interact with particular DNA sequence together with other TFs either through protein-protein interactions or through conformational changes of DNA[14]. In most cases, the recognition sites of the heterodimer TFs are quite different from each individual TF's binding sites based on our *in vitro* test[14], arising mainly from the overlap between individual TF recognition motifs. In some other cases, the dimeric motifs look like combination of both motifs, but the flanking space or the orientation of the motifs could be quite strict[14].

Because of diverse binding specificities of different types of TFs, the TFs are considered to be the core apparatus for specificity in gene regulation. They are responsible for recruiting other types of transcriptional regulators as well as RNA polymerase to the right positions of the genome, leading to the repression or induction of target gene expression. By providing specificity to gene regulation, TFs are thus essential in determining the cell identity. Ectopic expression of specific TFs can alter cellular identity. For instance, the epigenome and transcriptome of mouse fibroblast can be reprogrammed into that of the induced pluripotent stem cell (iPSCs), a cell type similar to the embryonic stem (ES) cell, by overexpressing a small set of TFs (**Figure 1.2**)[3,15,16]. The fibroblasts can also be directly converted to muscle cells by introducing only one TF, MyoD[2]. Therefore, introduction of a small set of TFs is sufficient to change the cell fate, indicating that the core transcriptional regulatory network responsible for establishing the whole transcriptional program for a particular cell identity is mainly composed of a small set of TFs.

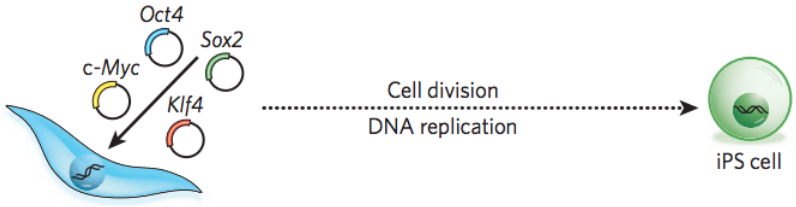

Figure 1.2 Induction of pluripotent stem cells by introducing TFs

The somatic cells are reprogrammed to the ES cell-like induced pluripotent stem cells (iPSCs) through introducing four TFs (*Oct4, Sox2, Klf4* and *c-Myc*), which have been shown to play important roles in maintaining pluripotency of the ES cells. Figure is adapted from [17].

1.1.2 Post-translational modifications of TFs

Like other kinds of proteins, such as enzymes, the TFs are also subject to post-translational modifications, resulting in changes in either their activities or cellular locations or interactions with other proteins. In the past few years, it has become apparent that different types of post-translational modifications, such as phosphorylation, ubiquitination, sumoylation and acetylation affect various aspects of TF functionality. The influence of these modifications is highly dependent on the

protein sequence of TFs, and adds an additional layer of complexity in regulation of gene expression by TFs.

Protein phosphorylation and dephosphorylation are carried out by protein kinases and phosphatases, respectively that are usually activated by multiple extracellular signaling pathways. Phosphorylation of TFs regulates their functions through different mechanisms: for instance by controlling the localization of the TFs, by modulating interactions with other factors, and by regulating binding activity with DNA. For instance, the SMAD family TFs, SMAD2 and SMAD3 are activated by the members of the transforming growth factor-β (TGF-β) family cytokines, leading to the phosphorylation of the serine residues at the carboxyl end of SMAD proteins[18]. The phosphorylated SMAD proteins can then associate with the common SMAD binding partner SMAD4 and be translocated to the nucleus to regulate gene expression. In contrast, the DNA binding activity of bZIP family TF c-JUN is inhibited when the COOH-terminal residues (threonine 231, serine 243 and serine 249) are phosphorylated[19]. Phosphorylation of its NH2-terminal transactivation domain on the other hand causes conformational change and dephosphorylation of the COOH-terminal residues, leading to increased DNA binding[19,20].

Ubiquitination is a process of addition of ubiquitin, a 76 amino acid peptide to the substrate proteins. The ubiquitination of proteins is usually associated with protein degradation via proteasome[21], but in recent years emerging data from yeast has revealed that ubiquitination can also regulate TF activity in a proteolysis-independent manner. For example, Kaiser *et al.* have demonstrated that in the yeast ubiquitination of TF MET4 lead to inactivation rather than degradation of the protein[22]. Moreover, the monoubiquitination of TF GAL4 was reported to stabilize transcription factor-DNA interactions in yeast[23]. However, regulation of TF activity through ubiquitination of proteins in proteolysis-independent way has not been detected in the mammalian cells.

Sumoylation refers to covalent attachment of SUMO, a small polypeptide, to the lysine residues of the substrate proteins[24,25]. Although not well studied yet, it has been shown to affect TF activity, stability and localization. The activities of TFs such as SP3[26], MYB[27] and CEBP[28] family TFs are all influenced by the SUMO-modification. Furthermore, because the lysine residues of proteins can also be post-translationally modified by phosphorylation, ubiquitination and acetylation, sumoylation can alter the transcription factor activities by competing with other types of modifications at the target lysines. For example sumoylation of the IκBα- the inhibitor of TF NF-κB at lysine residues K21 can stabilize the IκBα protein by blocking ubiquitination of the same residue, leading to inhibition of NF-κB activity[29].

The protein acetylation is controlled by two types of enzymes, lysine acetyltransferases (KATs) responsible for transferring the acetyl groups to lysine residues of the proteins and lysine deacetylases (KDACs) responsible for removing the acetyl groups from the acetylated lysine residues[30]. Different kinds of TFs are subject to acetylation, leading to their activation/inactivation or changes in their cellular localization. For instance, acetylation of the three lysine residues (Lys-242, Lys-245,

and Lys-262) within the DNA-binding domain of FOXO1 TF diminishes its ability to interact with DNA[31,32], whereas the lysine residues (Lys-91, Lys-94, and Lys-136) located within the transactivation domain of TF CREB when acetylated by CBP/p300, leads to increased transactivation activity[33]. In most cases, acetylation of lysine residues within the DNA-binding domains of TFs attenuates the interactions between TFs and DNA. However, there is one exception, acetylation of lysine residues within the DBD of GATA1 by CBP/p300 leads to enhanced DNA binding activity both *in vitro* and *in vivo*[34].

In addition to the above mentioned modifications, the TFs are also subject to other types of modifications including methylation and glycosylation. Methylation usually occurs at the arginine and lysine residues of the TFs, and the deregulation of TF methylation is frequently linked with diseases such as cancer[35,36]. The addition of O-linked beta-N-acetylglucosamine (O-GlcNAc) to the serine or threonine residues leads to the glycosylation of TFs. O-GlcNAc modification modulates functions of TFs in different ways, such as affecting their DNA-binding activities, localization and stability[36,37].

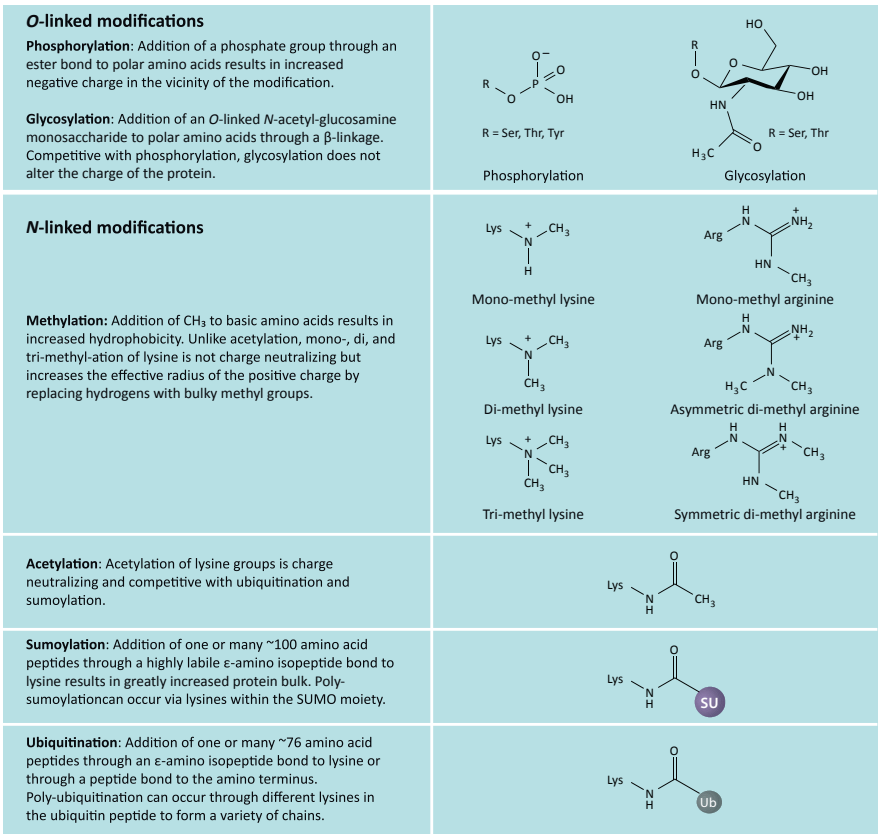

Figure 1.3 Types of post-translational modifications on TFs

The six major post-translational modifications occur on the TFs. They can be divided into two groups according to the atoms that are modified: the "O-linked modifications" including phosphorylation and glycosylation affect the oxygen atoms of the proteins,; the "N-linked modifications" including methylation, acetylation, sumoylation and ubiquitination affect the nitrogen atoms of the proteins,. Figure is adapted from [36].

In summary, the activity of TFs is highly dependent on the post-translational modifications that change the cellular localization, DNA binding activity, and transactivation activity of TFs. Therefore, it is quite important to measure the binding activity, rather than only the abundance of TFs in the cell nucleus to understand the mechanism of transcriptional regulation more thoroughly.

1.1.3 Technologies to study binding specificities of TFs

There are a variety of technologies utilized to study the TF binding with DNA both *in vivo* and *in vitro*. The *in vivo* methodology focuses on the interactions between the proteins and genomic DNA in cells; the *in vitro* methodology studies the interactions between the proteins and the *in vitro* synthetized DNA molecules. Compared with the *in vivo* assays, the *in vitro* assays are performed in simplified systems without the interference from other nuclear components. On the other hand, the results from *in vivo* assays represent the combined effect of variability in both the TF features as well as the DNA features. In different cell types or even in different loci of the genome from the same cell type, the status of the DNA with respect to its modification and TF accessibility is quite variable. Moreover, the interactions between TFs and DNA *in vivo* are also influenced by other factors such as non-coding RNA. Thus, the *in vivo* methodology to study TF binding is advantageous as this combined information is directly related to the binding events happening in cells, which is crucial for us to understand the mechanism of transcriptional regulation in cells.

The typical *in vivo* method to study TF binding in the genome is the chromatin immunoprecipitation (ChIP) assay, where the proteins bound to the genome are first covalently cross-linked to the DNA, followed by fragmentation of the chromatin and purification of desired proteins as well as cross-linked DNA fragments using specific antibodies[38]. The ChIP assay has been instrumental in identifying the interactions between DNA fragments from specific genomic regions and a wide range of nuclear proteins including TFs, modified histones, and chromatin remodelers. Moreover, ChIP assay followed by high-throughput sequencing technologies has further extended the scope of these experiments to genome-wide identification of TF binding locations[38,39], resulting in thousands to millions of binding sites identified throughout the genome. The TF binding specificity can potentially be determined by analyzing the genome-wide binding sites with the help of various kinds of computational algorithms such as MEME[40], Weeder[41], MDScan[42], and WebMOTIFS[43]. However, *de novo* motif discovery based on the ChIP-seq data for different TFs has indicated that it does not work so well for most TFs, partially because the DNA fragments pulled down together with the tested TFs are one or two order of magnitudes larger than a typical motif resulting in inherent noise. In order to identify the TF binding sites at a higher resolution, different groups have modified the ChIP technology and established novel technologies such as ChIP-exo[44,45] and ChIP-nexus[46] to achieve higher resolution. Both of these methods utilize the lambda exonuclease to degrade the protein-bound DNA in

a 5'-to-3' direction until it is blocked by the cross-linked proteins, leading to the enrichment of precise genomic loci bound by the TFs.

Compared with the *in vivo* assays, the *in vitro* assays originated much earlier and are usually applied to study the TF binding specificities *de novo*. DNA footprinting is one of the classical techniques to examine the binding of proteins to specific DNA sequences. It exploits the fact that when a TF is bound to DNA with a certain affinity, the DNA is then protected from degradation by nucleases. After digestion with nucleases, the bound and unbound DNA oligos can be separated on a polyacrylamide gel[47,48]. The nitrocellulose filter binding assay, developed in the early years of molecular biology, is another classical method to study the protein-DNA interactions. This assay relies on the specific chemical property of the nitrocellulose membrane to retain the proteins together with the bound DNA oligos and at the same time have quite low binding affinity with the free double-stranded DNA[49-51]. This is used to enrich and analyze the double-stranded DNA bound by the proteins. In comparison to these two assays described above, the electrophoretic mobility shift assay (EMSA) is a technically simple and relatively rapid method to detect protein-DNA interactions. The principle of this method is that the electrophoretic mobility of a complex composed of nucleic acid and protein is less than that of the free nucleic acid, resulting in the separation of bound and unbound DNA ligands[52-54]. Moreover, the EMSA assay can not only be utilized for qualitative purposes, but also provide quantitative information for determining binding affinities and kinetics[55].

Development of advanced technologies such as the microarray-based techniques and the massively parallel sequencing techniques have led scientists to combine these with the traditional methodologies to determine the TF binding specificity in a high throughput way. The protein binding microarrays (PBMs) introduced the use of DNA microarray technology to study the interactions between individual TFs and DNA (**Figure 1.4a**)[56-60]. The DNA microarray contains millions of double-stranded DNA oligos that have all the potential binding sites for all the TFs tested. The sequences of the DNA oligos either originate from the genome[59] or are completely randomized[58]. After incubating the microarray with tested protein carrying an epitope tag, the PBM is washed to remove any nonspecifically bound protein and then labeled with a fluorophore-conjugated antibody specific for the epitope tag; the fluorescence intensity for each individual spot on the microarray represents the binding affinity of the specific DNA sequence to the tested protein. Additionally, the protein microarray technology has also been utilized to identify the interactions among various proteins that recognize a particular sequence of interest, expanding our knowledge about protein-DNA interactions[61-63]. Apart from the microarray based technologies, another strategy used for studying protein-DNA interactions is the yeast one-hybrid (Y1H) system[64,65], which is conceptually similar to the classical yeast two-hybrid (Y2H) system used for detecting protein-protein interactions *in vivo*[66,67]. In the Y1H system, a DNA sequence (the DNA bait) is cloned upstream of a reporter gene and integrated into the yeast genome; meanwhile, the hybrid protein is generated by fusion of the prey protein to a transcription activation domain (**Figure 1.4b**). The expression level of the reporter gene

in the yeast cell can reveal how strongly the prey protein interacts with the DNA bait. By generating libraries of DNA constructs with the Gateway cloning system, the throughput of the study can be greatly enhanced[64]. A similar system but based on bacteria rather than yeast has also been established to study the DNA-binding specificities of Homeobox TFs[68].

The systematic evolution of ligands by exponential enrichment (SELEX) technology was first introduced in 1990 to detect specific RNA ligands that have the highest binding affinity to the bacteriophage T4 DNA Polymerase from a population of RNA ligands with random sequences[69]. The mechanism of the SELEX technology is similar with the process of evolution: during multiple rounds of enrichment, the best binding ligands are selected from a variety of random sequences, resulting in the exponential increase of the selected ligands. The SELEX protocol was also applied to study the binding specificity of TF Lrp to DNA sequences[70]. Later on, original work from our lab significantly increased the throughput of the SELEX assay by utilizing the automation system in the SELEX assay and combining it with the massively parallel sequencing system. The resulting high-throughput SELEX (HT-SELEX) assay can be performed within a week and characterizes, in parallel, binding specificities of hundreds of TFs (**Figure 1.4c**)[71,72]. In the HT-SELEX assay, the TFs are expressed in the bacteria with a specific epitope tag and purified in 96-well or 384-well plate format. With the help of the automation system, the TFs are incubated with double-stranded DNA composed of random sequences in the middle and the amplification adapters on both ends, following by washing for dozens of times and then purified together with the DNA still bound to the proteins; the enriched DNA is directly amplified by polymerase chain reaction (PCR) and used as a new DNA pool for another round of selection; finally, after several cycles of enrichment, the DNA ligands with the highest binding affinity for a particular TF are most selected and become dominating in the population of DNA pool. The HT-SELEX assay allows for systematic study of TFs binding specificities, and by means of this technology, we have applied the assay to more complicated studies, such as effect of TFs dimerization on interacting with DNA[14], and the systematic studies on interactions between TFs and methylated DNA[73].

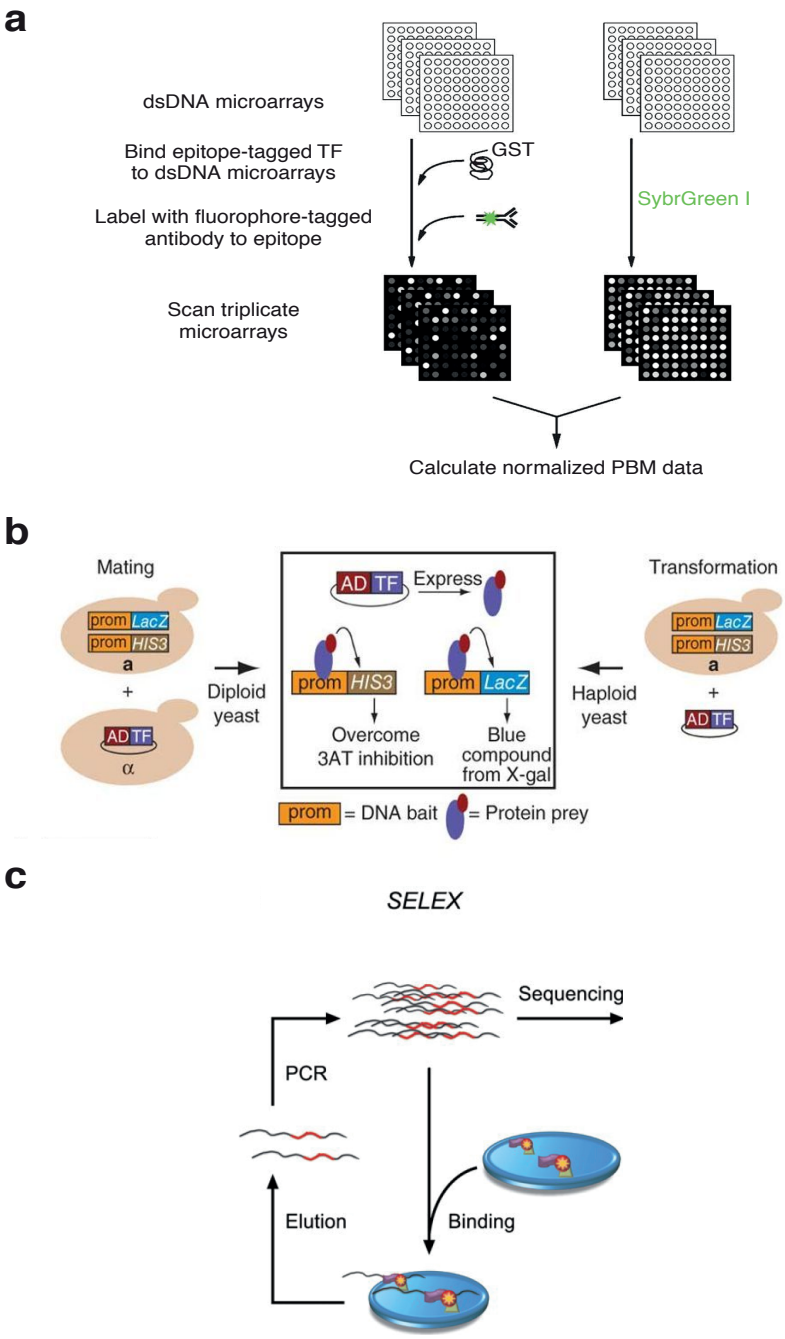

Figure 1.4 High-throughput technologies to study the TF binding specificity

a) Schematic of the PBM experiment. The double stranded DNA (dsDNA) microarray containing all the potential binding sites for all TFs tested is incubated with particular tested TFs carrying an epitope tag (GST), followed by washing and labeling with fluorophore-conjugated antibody targeting the epitope tag. After normalizing to the control microarrays, the fluorescence intensity for each spot reveals the binding affinity of the tested TF to the DNA sequence. To normalize the signals in each spot, a separate microarray from the same print is probed with SybrGreen I. Figure is adapted from [60].

b) Schematic of the Y1H assay through mating and transformation. prom indicates the promoter, also known as the DNA bait; the AD is the Gal4 transcription activation domain; TF indicates the tested transcription factor. The HIS3 and the LacZ are two independent reporter genes, reflecting the ability of the yeast to overcome 3AT inhibition and/or turn blue, respectively. If both phenotypes are observed in one cell, it means that the interaction between TF and DNA bait is very strong. Figure is adapted from [65].

c) Schematic of the SELEX assay followed by massively parallel sequencing. A dsDNA pool with randomized sequences is incubated with the immobilized TF; after stringent washing process, the DNA ligands still bound to the specific TF are eluted and amplified; the amplified DNA are further used for selection during several cycles followed by sequencing. Figure is adapted from [72].

1.1.4 Structured transcriptional regulatory network

Similar with other real-world complex systems we are familiar with, such as the World Wide Web (WWW)[74,75], the transcriptional regulatory network is interpreted as a scale-free network with hierarchical organization[76]. As a standard scale-free network, the basic feature is that the degree distribution of the network follows a power-law, where the probability $P(k)$ that a node in the network interacts with k other nodes decays as a power law($P(k) \sim k^{-\gamma}$)[77], indicating that the network contains numerous nodes with only a few links and a small number of nodes with a huge amount of links. Those nodes with a huge amount of links are the hubs of the network, and establish the core of the whole network. As for the transcriptional regulatory network, thousands of TFs are expressed in the cells, but only a small number of them act as the hubs to play dominant roles to interact with other TFs and cofactors to set up the transcriptome profile for the specific cell identity; these dominant TFs are considered as the master regulators of the cell fate.

In addition, the scale-free networks have several other properties. Firstly, it can be freely expanded, indicating that new TFs can be easily supplemented to the existing transcriptional regulatory network. This is a very important feature as it speeds up the process of evolution by efficiently adding new factors to the original network. Another important property is called "preferential attachment", meaning that every new node would preferably connect with already well-connected nodes, the hubs. Therefore, it is likely that the most important TFs (hubs) of the network remain the hubs of the network in new species during evolution, resulting in the conservation of core transcriptional regulatory network during evolution. More importantly, the scale-free network is especially robust against accidental failures, which makes the organisms much more adaptive to different kinds of perturbations from the environment, and also makes the developmental program of multicellular organisms to be smoothly executed under fluctuating conditions.

Apart from that, the hierarchical structure of the network indicates that the network can be divided into different groups that can be further subdivided into smaller groups over multiple scales[78], implying that different types of factors including TFs are clustered into different groups to regulate gene expression hierarchically.

All in all, properties of the transcriptional regulatory network indicate that a small number of TFs act at the highest hierarchy to regulate gene expression for specific cell identity. Therefore, in order to understand the mechanism of transcriptional regulation in different types of cells, it is crucial to first determine the core TFs responsible for a particular cell identity. These TFs are expected to possess the highest binding activity on the genome in a given cell type to recruit relevant cofactors in order to induce or repress gene expression.

There was no method to detect the DNA-binding activities of all TFs inside the cells systematically. Existing technologies such as RNA-seq or proteomics only analyze the RNA or protein levels[5,6]; the ChIP-seq technique does measure the TF activities to

some extent, but it cannot compare activity levels between different TFs owing to the diverse binding affinities of antibodies to specific TFs[71,79,80]. In order to measure the DNA-binding activities of TFs systematically and identify the most active ones, in this project we aimed at establishing a novel massively parallel protein activity assay to measure the DNA-binding activity of all TFs from cell extract.

1.2 NUCLEOSOME OCCUPANCY & CHROMATIN ACCESSIBILITY

One of the most important features of eukaryotic genome is packaging of DNA into nucleosomes and chromatin. The nucleosome is the fundamental unit of chromatin, and is composed of a histone octamer core spiraled around by about 147 base pair (bp) double stranded DNA (**Figure 1.5a**)[81,82].The nucleosomes are arranged as a linear array along the DNA polymer like "beads on a string" (**Figure 1.5b**)[83,84]. The formation of nucleosomes not only allows the genome to be folded into chromatin and compacted by thousands of times to fit in the tiny nucleus, but also sharply increase the complexity of gene regulation in eukaryotes compared with the prokaryotes. The occupancy of nucleosomes in the genome is closely related to gene expression for specific cell identity, as the accessibility of the regulatory elements is directly related to gene transcription[85].

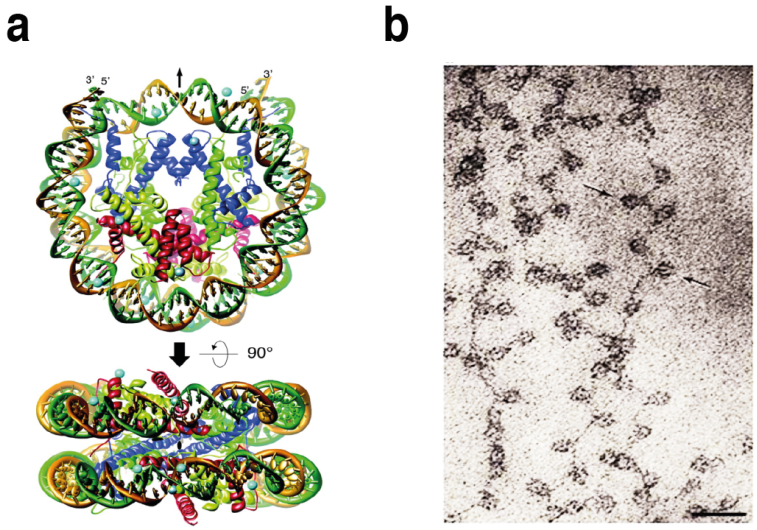

Figure 1.5 Nucleosome core particle (NCP) octamer

a) The typical crystal structure of a human nucleosome core particle (NCP). The arrow above the upper ribbon model indicates a pseudo 2-fold axis passing through the center of the particle; the bottom figure shows the axial view of the particle. In the core of the model, different histones are shown in different colors: yellow for H2A, red for H2B, blue for H3 and green for H4. The brown and green ribbon traces around the core histones display the phosphodiester backbones of the DNA for both chains. The Mn^{2+} and Cl^{-} are depicted as cyan and silver balls, respectively. Figure is adapted from [81].

b) The electron micrograph of chromatin under low ionic-strength solution. The two arrows indicate two nucleosomes along the DNA polymer like the "beads on a string". Scale bar, 30 nm. Figure is adapted from [84].

1.2.1 Nucleosome occupancy regulates transcription

The occupancy of nucleosomes regulates transcription in different ways. The most direct way is to inhibit transcription initiation by blocking the assembly of transcriptional machinery in the core promoters. It has been proved that the presence of nucleosomes impedes transcription *in vitro*[86,87], and the nucleosome loss is able to expedite the transcription initiation *in vivo*[88,89]. Genome-wide mapping of nucleosome positioning in species from yeast[90-92] to humans[93,94] has indicated that active promoters are depleted of nucleosomes, restricting the assembly of PIC to specific genomic regions for expression of particular genes. Moreover, the positioning of nucleosomes in the genome also allows for direct entry of transcriptional machinery to the promoter rather than the middle of the gene, which significantly decreases the transcriptional noise.

In addition to blocking the assembly of transcriptional machinery to initiate transcription, the nucleosome occupancy also prevents other DNA binding factors, such as TFs from binding to the genomic DNA. For instance, Yuan and colleagues have shown that most loci (over 87%) bound by TFs were within nucleosome-free regions in the budding yeast[95], implying that the interactions between TFs and genomic DNA are usually blocked by nucleosomes. Based on this phenomenon, a "site exposure" model was also proposed to elucidate the competition between TFs and histones on binding the genomic DNA, assuming that DNA on the surface of nucleosomes is transiently exposed due to thermal fluctuation[96,97]. Furthermore, original work from our lab provided more direct evidence that almost all TFs' binding events with DNA were blocked by nucleosomes[98].

Moreover, the occupancy of genomic DNA by nucleosomes can also facilitate access of epigenetic modifiers to the targeted regions to modify the DNA or histones, leading to condensation of the chromatin and repression of gene expression. For example, the DNA methylation at the regulatory elements is well known to be closely correlated with gene repression[99-101], and it has been indicated that the *de novo* DNA Methyltransferases DNMT3A/3B are selectively anchored to a subset of nucleosomal DNA, which only requires the intact nucleosomal structure rather than the presence of other chromatin-modifying enzymes or proteins for their recruitment[102].

Because the nucleosome occupancy regulates binding of transcriptional machinery as well as other DNA binding factors, it is expected to play the causal role in determining transcriptome profiles in specific cell identity.

In addition to regulating gene expression, the occupancy of nucleosomes in the genome also plays a critical role in other kinds of DNA related processes, such as DNA replication and DNA repair[103-105]. Generally, all these processes need the proteins to get access to the genomic DNA, therefore packaging of genomic DNA into chromatin inhibits the DNA-dependent processes. Although DNA replication and repair are not directly related to the transcription of genes, they also have effect on the cell identity by altering the cell cycle length[106] or mutating the crucial genes or regulatory elements.

1.2.2 Technologies to study chromatin accessibility

It is clear that chromatin accessibility is very important for gene expression and cell fate determination. Therefore, identification of the accessible genomic regions in specific cell identity is necessary for us to understand the mechanism of cell fate determination. In combination with massively parallel sequencing technologies, different types of methods have been utilized to determine the genome-wide chromatin accessibility in particular cell or tissue types. One classical method is to use low concentration of nuclease enzyme Dnase I to cleave the genomic regions that are depleted of nucleosomes (**Figure 1.6**); the DNase I is so active that high concentration of the enzyme can even digest DNA exposed in the nucleosome when wrapping around histones[107]. The regions with preferential digestion of DNase I are referred to as DNase I hypersensitive sites (DHSs), including all the active cis-regulatory elements, such as enhancers and promoters, which are important for regulation of particular genes[108-111]. Furthermore, because the binding of TFs affects the intensity of DNase I cleavage and generates footprints of the TFs in the genome, the DHSs can also be applied to study the TF occupancy in a qualitative and quantitative manner[112]. Apart from DNase I, other types of nucleases are also used to study the DNA accessibility. One of the most commonly used one is the micrococcal nuclease (MNase), which is an endo-exonuclease that preferentially digests single-stranded nucleic acids and also has activity against double-stranded DNA. The activity of MNase is weaker than the DNase I as digestion by MNase can be obstructed by not only nucleosomes, but also other types of DNA binding proteins, therefore some DHSs bound by factors such as TFs are excluded from digestion by MNase (**Figure 1.6**). Owing to its weak activity, MNase has been applied to other studies beyond the nucleosome analysis. For example, paired-end MNase-seq has been used to map the distribution of paused RNA polymerase II in *Drosophila* S2 cells[113]; in addition, the MNase digestion was also conjugated with the ChIP technology to identify the binding sites of chromatin structure remodeling (RSC) complex in yeast[114].

Apart from nuclease digestion, the accessible genomic regions can also be detected with the formaldehyde-assisted isolation of regulatory elements (FAIRE) technology, which entails formaldehyde fixation of chromatin and subsequent separation of protein-free DNA by phenol–chloroform extraction[115]. Compared with the nuclease digestion, the FAIRE technology does not require the isolation of nuclei, and it overcomes the cleavage bias caused by the nucleases[116-118]. However, there appears one major limitation of FAIRE that outweighs all the benefits: the signal-to-noise ratio is relatively lower than other chromatin accessibility assays so that only strong signals could be informative (**Figure 1.6**).

More recently, a novel technology named ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) has been introduced to study chromatin accessibility in quite rare biological samples. The principle of the technique is to fragment and amplify genomic DNA within the open chromatin regions by using the

hyperactive Tn5 transposase loaded with sequencing adapters and proceed to next generation sequencing[119]. The ATAC-seq technique is quite sensitive as in order to get similar results from ATAC-seq data, three to five orders of magnitude more cells are required in DNase-seq[119]. Besides, since there is no size-selection step in the ATAC-seq protocol, it can also be applied to identify nucleosome positioning and the accessible regions simultaneously based on the size of the amplicons, as only sequences longer than 147 bp are occupied by nucleosomes, and sequences shorter than 147 bp are from accessible regions. Because of the high sensitivity, the ATAC-seq was further utilized to study chromatin accessibility in individual cells[120], revealing great potential of this highly attractive technology in the single cell studies.

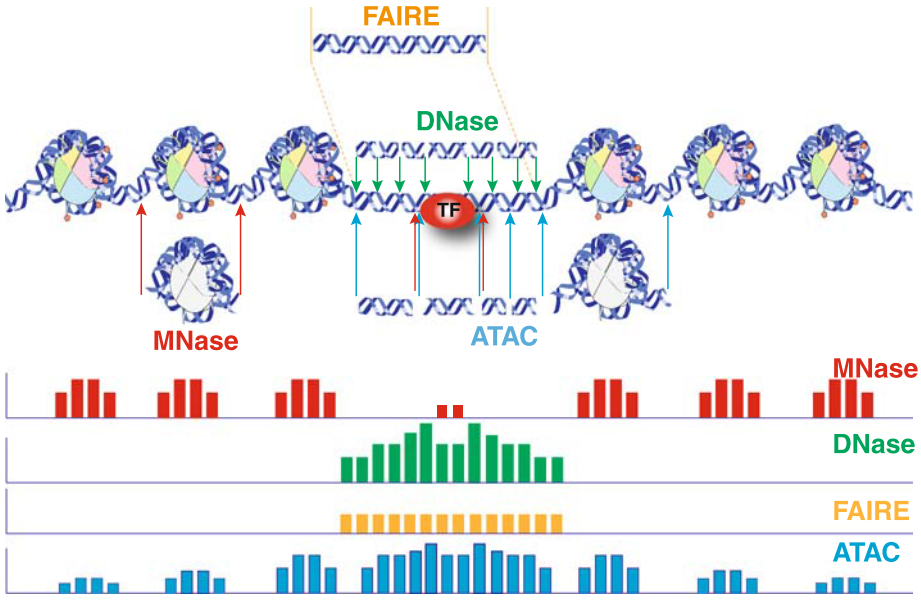

Figure 1.6 Schematic diagram of current technologies for studying chromatin accessibility

Different technologies to study chromatin accessibility are displayed and compared by taking one genomic region as example. The colored arrows indicate the boundary of the DNA extracted from different techniques; the colors of the arrows represent different techniques, which are the same as those shown in the histograms below. The histograms indicate data signals obtained from each assay across the entire region. Figure is modified from [121].

1.2.3 Determinants of chromatin accessibility

Since the chromatin landscape of the genome determines the transcriptome profile for specific cell identity, we need to find out the determinants of chromatin accessibility in order to decipher the determinants of gene regulation. Overall, the chromatin accessibility is generally determined by DNA sequences (*cis*-acting factors) and by epigenetic regulators such as chromatin remodelers, transcription factors as well as the epigenetic modifiers (*trans*-acting factors).

Sequence features. Although the histone octamer has much lower sequence binding preference than the sequence specific TFs, the flexibility of DNA sequences plays an important role in interacting with the histone octamers as the energy required to bend different DNA sequences to wrap around a small octamer of proteins varies a lot, resulting in variation in the stability of the formed nucleosomes[122]. Unlike the binding motifs for different types of TFs, there are favorable and unfavorable binding patterns resulting from the local bendability of the sequence for nucleosomes to bind to the DNA. For example, the poly-dA stretches which are intrinsically stiff and superabundant in the eukaryotic genomes relative to the prokaryotic genomes[123] have been shown to disfavor nucleosome formation both *in vitro*[124-126] and *in vivo*[95,126,127]. Conversely, the periodic A/T dinucleotides (AA, TT or TA) spaced at 10 bp intervals reveal higher binding affinity for histone octamer than random sequences; this pattern has also been observed with statistically high significance in different eukaryotic genomes[128-130].

Moreover, the importance of sequence features in determining chromatin accessibility has also been proved by predicting nucleosome positioning throughout the yeast genome with quite high accuracy based only on the data from *in vitro* reconstitution of genomic DNA into histone octamers[92]. In addition to the yeast, genome-wide analyses from higher multicellular organisms such as flies[128,131] and humans[132] also demonstrate that the sequence information of DNA plays important roles in determining the chromatin accessibility. However, the power for predicting the nucleosome positioning in flies[128,132] and humans[94,132] is much lower than that in the yeast, implying that apart from the sequence features, *trans*-acting factors, such as chromatin remodelers, TFs, and epigenetic enzymes play more important roles in setting up the whole genome accessibility in those species.

It should be noted that because in the multicellular organisms, all cells share the same genetic background, meaning that the sequence features of the genome are identical for different cell identities, the cell specific chromatin landscape is thus determined by the *trans*-acting factors.

***Trans*-acting factors.** The *trans*-acting factors, including TFs, epigenetic modifiers, and chromatin remodelers, act as another major determinant of chromatin accessibility.

As mentioned earlier, TFs usually compete with nucleosomes to interact with DNA, leading to the ejection of histones out of the DNA at specific loci or *vice versa*:

nucleosomes prevent TF binding. However, in rare cases, the TFs can also promote nucleosome assembly via their intrinsic nucleosome-assembly activities[133] or by recruiting other factors such as polycomb repressive complex 2 (PRC2) to specific regions[134].

The epigenetic modifiers are proteins responsible for epigenetic changes, including modifications of DNA and histones. In eukaryotes, DNA modifications mainly include methylation and demethylation on the 5th carbon atom of the DNA base cytosine[135], as well as the intermediate of these two major states, such as hydroxymethylation[136]. Nowadays, only the methylation on 5th carbon atom of cytosine (5mC) is considered to be heritable and involved in transcriptional regulation; the functions of the other intermediate modifications are still unclear. Methylation of cytosine is catalyzed by the DNA methyltransferases (DNMTs). In vertebrates, there are four members of the DNMTs family, including DNMT1, DNMT3A, DNMT3B and DNMT3L. DNMT3L almost has no intrinsic enzymatic activity because of the lack of most of the C-terminal catalytic domain compared to other members[137]. The other three DNMTs are the active methyltransferases responsible for the methylation patterns of DNA throughout the genome. DNMT1 is essential for the maintenance of the DNA methylation during DNA replication[138], and DNMT3A/3B serve as *de novo* DNA methyltransferases[139]. As mentioned earlier, DNA methylation is highly associated with gene repression. One of the reasons is that the methylated cytosine would promote formation of heterochromatin, leading to the condensation of the chromatin[140,141].

Demethylation of cytosine occurs in vertebrates in two different ways: the passive and the active way. In the passive way, methylated DNA is diluted by successive DNA replication without maintenance of the methylation pattern through DNMT1; for example, during the second and third cell cycle after fertilization, the mouse maternal genome is gradually demethylated owing to the loss of DNMT1 activity[142]. In addition, the ten-eleven translocation (TET) proteins discovered in almost all the metazoans are responsible for the active demethylation of cytosine by oxidizing the methyl group to hydroxymethyl group[143,144]. It is still questionable whether hydroxymethylation of cytosine (5hmC) has any regulatory role in gene expression, but steadily increasing evidence shows enrichment of 5hmC at specific genomic regions in particular cell types, including the ES cells[145,146] and neurons in the central nervous system[147].

Histone residues, especially the tails, are subject to more than 100 different posttranslational modifications, including methylation, acetylation, and phosphorylation; some of these modifications have shown significant correlation with the chromatin accessibility and transcriptional processes. For example, high level of histone 3 lysine 4 trimethylation (H3K4me3) around the transcription start site and histone 3 lysine 36 trimethylation (H3K36me3) at the 3' open reading frame (ORF) indicate active transcription of genes[148,149]; enrichment of acetylation of H3 lysine 9 (H3K9ac) at the promoter or other regulatory regions displays accessible chromatins and gene activation[150]; whereas methylation of H3 lysine 9 (H3K9me) is closely

associated with heterochromatin formation[151,152] and DNA methylation[153,154], which leads to transcriptional repression.

Because of the close correlation between the epigenetic modifications and chromatin accessibility, the epigenetic enzymes responsible for these modifications play important roles in determining the chromatin accessibility. Apart from that, some epigenetic modifiers are even exclusively expressed in particular cell types, suggesting the essential roles of these factors in the specific cell types. For example, the DNA demethylase Tet3 is specifically expressed in mouse zygotes and oocytes[155,156], implying the crucial roles of Tet3 in specification of mammalian gametes; decreased expression and enzymatic activities of Tet family proteins, which lead to hypermethylation of genomic DNA, also play important roles in development of cancers, such as acute myeloid leukemia[157] and melanoma[158]. In addition, many histone modifiers are also expressed only in certain cell types, suggesting the importance of those factors in cell fate determination. For example H3K36 demethylase Kdm2b[159], and the common component of the H3K4 methyltransferase complex Wdr5[160] are specifically expressed in ES cells. Moreover, these histone modifiers are able to enhance the efficiency of somatic cell reprogramming to the iPSCs with the help of other factors[160,161], indicating the crucial roles they play in maintaining pluripotency of the ES cells.

Chromatin remodelers are usually recruited by TFs or modified histone residues at specific loci and utilize the energy generated from ATP hydrolysis to change the nucleosomes' architecture through incorporating or ejecting histone octamers, sliding nucleosomes and altering nucleosome composition by histone exchange. The chromatin remodelers are composed of several subunits, and each subunit has distinct functions with respect to the remodeling of chromatin structure. Each chromatin-remodeling complex contains one ATPase subunit in addition to other regulatory factors mediating protein-protein interactions and protein-chromatin interactions[162,163]. According to the unique domains included in the ATPase subunit, the chromatin remodelers are divided into four families: SWI/SNF (switch/sucrose non-fermentable), CHD (chromodomain-helicase DNA-binding), ISWI (imitation switch), and INO80 (inositol-requiring 80)[162,163]. The mechanism of how different chromatin remodelers cooperate with histone modifications as well as other transcriptional regulators to alter the chromatin architecture is not clearly elucidated. Recently, researchers have determined the binding of remodelers to individual nucleosomes in the yeast genome in a high-throughput way, and found that different remodelers exhibit distinct and clear distributions with respect to transcribed genes[164], indicating key roles of these remodelers in transcriptional regulation. Furthermore, there is clear evidence showing that the sequence-specific TFs play crucial roles in targeting the chromatin remodelers (the ISWI family complex) to specific genomic loci, leading to the disruption of nucleosomes in those regions[165].

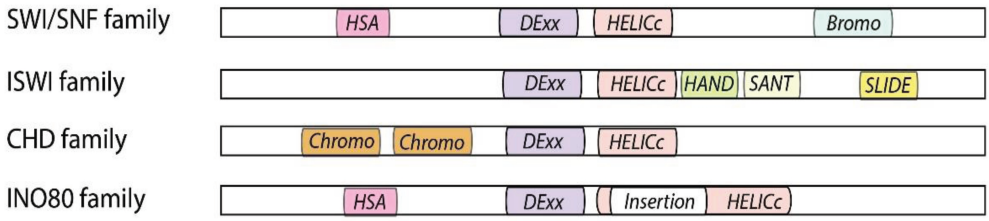

Figure 1.7 Classification of chromatin remodelers by the ATPase subunit

All the chromatin remodelers are categorized into four groups based on different types of domains included in the ATPase subunit. The DExx (purple) and HELICc (pink) are the two basic components of the ATPase domain. The INO80 family is distinguished from the other three families as it contains a long insertion (white) inside the HELICc domain. Other chromatin remodelers are further divided by distinct combinations of flanking domains: Bromo domain (light blue) and HSA (helicase-SANT) domain (red) for SWI/SNF family, SANT-SLIDE module (yellow) for ISWI family, tandem chromo domains (brown) for the CHD family. Figure is adapted from [162].

Although there are different types of *trans*-acting factors that determine the accessibility of the whole genome for specific cell identity, it is possible that the TFs could be the dominant ones by first targeting specific loci due to their DNA binding specificities, and then recruiting other regulators to induce local conformational changes of the chromatin. In this project, I have correlated the chromatin accessibility data with the TFs binding activities from the same cell types to test the hypothesis that the TFs, especially the most active ones play dominant roles in determining chromatin landscape for specific cell identity due to their DNA-binding activities.

1.3 PIONEER TRANSCRIPTION FACTORS

Although TFs are considered to initiate chromatin structural changes within specific loci, the ability to compete with histones for DNA binding differs a lot among individual TFs. Specifically, there are a group of TFs that have much higher binding affinity with nucleosomal DNA than others. These TFs serve as the pioneer factors to target the compacted genomic regions, and then recruit other TFs as well as cofactors to remodel the chromatin structures locally. The pioneer TFs are regarded as the predominant factors that initiate regulatory events within silent chromatin regions during transition of the cell identities.

1.3.1 Interactions between pioneer TFs and nucleosomes

The initial step in opening the condensed chromatin by the pioneer factors is their binding to the DNA target sites embedded in the chromatin. Pioneer TFs, unlike other TFs, have the special ability to interact with nucleosomal DNA. For instance, an important feature of the members of the classical pioneer TFs FOXA family is that part of their DNA-binding domains highly resemble that of the linker histone H1, which possess the "winged helix" motif that contacts the minor groove of the DNA along the long axis, facilitating binding of core histones on the other side of DNA[166,167]. In addition, both *in vitro*[168] and *in vivo*[169] experiments have shown that the FOXA TFs only require their DNA-binding domain to bind the target sequences on a nucleosome core particle. Another demonstration of interactions between pioneer TFs and nucleosomes comes from experiment where the classical Yamanaka factors (OCT4/POU5F1, SOX2, KLF4 and c-MYC) were introduced to reprogram the somatic cells to iPSCs[3]; it was shown that OCT4, SOX2 and KLF4 bound specific loci within condensed chromatin regions in the early stage of reprogramming, working as the pioneer TFs, whereas c-MYC were only able to bind open chromatin regions[170]. In order to determine the binding activities of different TFs on nucleosomes systematically, we have established a novel technology, nucleosome consecutive affinity-purification SELEX to systematically decipher the binding activity of 220 TFs from diverse structural families on nucleosomal DNA, greatly expanding the knowledge about the binding of TFs to nucleosomal DNA[171].

In addition, different kinds of modifications on histone residues, especially on histone 3 may also play important roles in affecting interactions between the nucleosomes and pioneer TFs. For instance, by comparing the epigenetic states of different genomic sites bound by FOXA1 in different cell types, Mathieu Lupien and colleagues found that FOXA1 as a pioneer factor was preferentially recruited to loci with high H3K4me1/2 and low H3K9me2 modifications[172]. There is no evidence indicating that there exists direct physical interaction between the FOXA1 protein and the H3K4me1/2 residue, and recruitment of FOXA1 pioneer factor might be influenced

by conformational changes of the chromatin induced by different epigenetic modifications.

All in all, in order to understand the principles of how pioneer TFs work to initiate the conformational changes of local chromatin, it is necessary to decipher how and to what extent different kinds of pioneer TFs interact with modified or unmodified nucleosomes in a systematic manner.

1.3.2 Pioneer TFs and development

Due to the special ability of pioneer TFs to interact with nucleosomal DNA with high affinity, they are expected to play more important roles in determining cell lineages during development than other TFs as the process of development involves extremely dynamic and accurate changes of the chromatin architecture in different loci.

It is already well known that the FOXA pioneer TFs are crucial for the early embryogenesis. Knocking out the *Foxa1* gene in mice leads to early postnatal lethality (P2-P14) due to hypoglycemia and defects in kidney function[173,174]. In addition, mice null for *Foxa2* die in the early embryonic stage (E9-E10) because of developmental defects in all three germ layers[175,176], whereas mice deficient for *Foxa3* have the mildest phenotype: they are viable but display more severe hypoglycemia after prolonged fast than the normal mice[177]. Apart from the FOXA family TFs, other pioneer TFs such as the GATA family proteins GATA4 and GATA6 also play important roles in early embryogenesis as mice lacking *Gata4* or *Gata6* die prior to the organ development[178,179].

Owing to the ability to bind the nucleosomal DNA with high affinity, the pioneer TFs are expected to spatiotemporally target specific loci in the genome during differentiation, and further recruit other interacting TFs as well as cofactors to execute modifications and conformational changes of the chromatin. For example, during the process of skeletal muscle differentiation, the pioneer factor PBX is responsible for targeting the muscle specific genes initially and recruiting other factors to induce gene expression[180]. It has also been proved that in early *Caenorhabditis elegans* foregut development, PHA-4, a homolog of *FOXA* genes in human, frequently binds promoters of gut-specific genes and recruits RNA polymerase II to enhance transcription of the genes[181].

Apart from actively initiating the local chromatin structural changes by facilitating recruitment of other factors to specific loci, the pioneer TFs can also passively reduce the number of subsequent binding events required for transcriptional activation at a later time owing to their presence at the specific regulatory regions. This passive model is further supported by the fact that the vast majority of promoters and enhancers, especially those related with lineage specific genes, always require cooperative binding of different TFs[182-184]. Furthermore, this cooperative binding model composed of different types of TFs also reveals the importance of other non-pioneer TFs during development.

In conclusion, the cell type specific pioneer TFs play crucial roles in targeting and opening up specific loci throughout the genome to induce transcription of the cell-type specific genes with the help of other non-pioneer TFs and co-regulators.

1.3.3 Pioneer TFs and tumorigenesis

Apart from playing important roles in initiating local chromatin structural changes during normal development, pioneer TFs are also considered to be important for significant epigenetic alterations of the genome as well as aberrant expression of genes during different stages of cancer development. For example, dysregulation of FOXA and GATA family TFs is closely related to a variety of hormone-dependent tumors, such as the oestrogen receptor-positive breast cancer and the androgen receptor-positive prostate cancer[185]; the increased activity of FOXA1 for prostate cancer is usually predictive of poor clinical outcome of the patients[186,187]. Moreover, upregulation of the pioneer factor SOX2 is highly correlated with different types of cancers including lung and esophageal squamous cell carcinomas[188]; SOX2 is also considered as a key factor maintaining the cancer stem cell population that drives tumor initiation and therapy resistance[189,190].

Alternatively cancers may also be caused by mutations of the binding sites of pioneer factors rather than alterations in their activity. For example, some of the recurrent somatic mutations in T-cell acute lymphoblastic leukemia (T-ALL) patients, are known to introduce high affinity binding sites for the pioneer TF MYB upstream of the *TAL1* oncogene resulting in *TAL1* overexpression[191]. Genome-wide association studies have also linked the genetic variants, such as single nucleotide polymorphisms (SNPs), to the development of cancer. The majority of the SNPs are located at the non-coding regions of the genome, implying that at least some of the mutations may alter binding of pioneer TFs such as FOXA1 in prostate cancer[192,193]. Moreover, epigenetic modifications rather than genetic variations could also alter pioneer TF binding as some specific modifications are favored by the TFs. For example, it has been demonstrated that multiple types of TFs, including pioneer TFs such as POU5F1 preferably bind methylated DNA[73]; it has also been reported that the progression of prostate cancer to an androgen-independent stage is highly correlated with the FOXA1 binding favored by increased H3K4 methylation at particular locus, promoting proliferation of castration-resistant prostate cancer cells, whereas removal of the H3K4 methylation in those cancer cells significantly reduces the FOXA1 binding and decreases cancer cell proliferation[194].

Because of the primary role of pioneer TFs in cancer development, they are regarded as attractive molecular targets and biomarkers for cancer therapy. Recently scientists have put considerable effort on designing new drugs to target the pioneer TFs for cancer treatment. For example, different types of small peptides have been designed to antagonize interactions between the pioneer factor PBX1 and HOX factors to promote apoptosis in melanoma, ovarian and lung cancer cells[195-197]. Although the

therapeutic approaches to antagonizing FOXA1 activity are still missing, the FOXA1 protein could act as a great prognostic biomarker for various kinds of cancers. For example, overexpression of FOXA1 is closely associated with metastasis in prostate cancer[186] as well as luminal subtype A breast cancer[198-200]. Co-expression of FOXA1 and GATA3 characterizes the luminal breast tumors and is a predictor of higher survival[200-202]. Targeting the pioneer TFs may have great potential in the future cancer treatments, however, a better understanding of how pioneer factors function to change the cell fate during initiation and development stages of cancer will be required in order to target them reliably for therapeutic breakthroughs.

Taken together, the pioneer TFs contribute significantly to cancer development either through mutations of their coding sequences, modulation of their expression or alteration of their genomic activities. They have thus great potential as therapeutic targets and biomarkers in cancer treatment.

2 AIMS OF THE STUDY

This study aimed at deciphering the most active and important transcription factors for specific cell identity. By establishing a novel technology to measure the global transcription factor activity in cells, we aimed to compare the DNA-binding activities of all TFs inside the cell nucleus, and decipher the dominant TFs that set up the whole transcriptional regulatory architecture in particular cell types.

More specifically, we divided the goal into three parts:

1) Establishing a novel method to measure the DNA-binding activities of all TFs in different cell types, and verifying the results by reprogramming of fibroblasts to the target cell identity.

2) Studying the correlation between binding activity of TFs and chromatin accessibility in the same cell identity to prove the dominant role of TFs in determining chromatin landscape for specific cell identity.

3) Determining the pioneer TFs for different cell identities by using nucleosomal DNA instead of naked DNA in the assay.

3 MATERIALS AND METHODS

3.1 MATERIALS

3.1.1 Reagents and commercial kits

The reagents and commercial kits used in the thesis are clearly indicated in the "**METHODS**" section of the thesis.

3.1.2 Cells lines

The mouse embryonic stem (ES) cells and MEF feeder cells were obtained from Karolinska Center for Transgene Technologies. The mouse ES cells were authenticated by production of germline chimeric mice, alkaline phosphatase staining and cell morphology.

The human fibroblast cell line CCD-1112Sk was purchased from ATCC (Cat no. CRL 2429); the human 293FT cell line was purchased from Thermo Scientific (Cat no. R70007); the *Drosophila* S2 cell line was purchased from Thermo Scientific (Cat no. R69007).

3.1.3 Animals

The mice used in this thesis were all adult male with the wild-type (C57BL/6J) genetic background.

3.2 METHODS

3.2.1 Cell culture & Protein extraction

The Mouse embryonic stem (ES) cells (C57BL/6J; from KCTT center at Karolinska Institutet) were first plated on 150 mm × 25 mm petri dishes in ES1+LIF medium with or without MEF feeder layers and changed to 2i+LIF medium after the cells attached to the plates; cells were split every two or three days and collected by trypsinization at 70-80% confluence, the MEF feeder cells were removed using differential adhesion method[203] for the final collection. The ES1+LIF medium is composed of knockout DMEM medium (Gibco, Cat no. 10829-018) supplemented by 15% fetal bovine serum (Sigma, Cat no. F7524), 2 mM L-Glutamine (Gibco, Cat no. 25030-024), 0.1 mM of each Non-Essential Amino Acids (Gibco, Cat no. 11140-035),

0.1 mM β-mercaptoethanol (Thermo Scientific, Cat no. 31350-010), 10 μg/ml Gentamicin (Thermo Scientific, Cat no. 15710-049), 10 mM HEPES (Gibco, Cat no. 15630-056) and 1000 U/ml Leukemia inhibitory factor (LIF, Millipore, Cat no. ESG1107). The 2i+LIF medium is composed of knockout DMEM medium with 20% KnockOut™ Serum Replacement (Gibco, Cat no. 10828-028), 2 mM L-Glutamine, 0.1 mM of each Non-Essential Amino Acids (Gibco, Cat no. 11140-035), 0.1 mM β-mercaptoethanol (Thermo Scientific, Cat no. 31350-010), 10 μg/ml Gentamicin (Thermo Scientific, Cat no. 15710-049), 1 μM MEK inhibitor PD0325901 (Miltenyi Biotec, Cat no. 130-103-923), 2 μM GSK-3α/β inhibitor BIO (Sigma, Cat no. B1686) and 1000 U/ml LIF.

The Human fibroblast cell line CCD-1112Sk (ATCC, Cat no. CRL 2429) and the highly transfectable 293FT cell line were both cultured in DMEM medium supplemented with 10% fetal bovine serum and antibiotics (100 units/ml of penicillin and 100 μg/ml streptomycin).

The S2 cells (Thermo Scientific, Cat no. R69007) from *Drosophila* were cultured in Schneider's *Drosophila* Medium (Thermo Scientific, Cat no. 21720024) at 27 °C without CO_2 and harvested by trypsinization.

The ES cells for neural differentiation were cultured with 2i medium supplemented with 2 μM retinoic acid (Sigma, Cat no. R2625-100MG) for 2 days; the ES cells for mesodermal differentiation were first cultured with 2i+LIF medium for 16 hours, and then changed to the mesodermal medium and cultured for 30 more hours; the control ES cells were cultured in 2i+LIF medium without feeder layers. The mesodermal medium (206 ml) includes: 100 ml IMDM (Thermo Scientific, Cat no. 12440053) supplemented with GlutaMAX (Thermo Scientific, Cat no. 31980030), 100 ml Ham's F-12 Nutrient Mix (Thermo Scientific, Cat no. 21765029), 2 ml N2 supplement (100×, Thermo Scientific, Cat no. 17502048), 4 ml B27 supplement (50×, Thermo Scientific, Cat no. 17504044), 0.5 mM ascorbic acid (Sigma, Cat no. A92902), 4.5×10^{-4} M monothioglycerol (Sigma, Cat no. M1753), 5 ng/ml VEGF (Thermo Scientific, Cat no. PHC9391), 8 ng/ml Activin A (Thermo Scientific, Cat no. PHG9014) and 0.5 ng/ml BMP4 (Thermo Scientific, Cat no. PHC9534).

The protein extraction was performed with "Subcellular Protein Fractionation Kit for Tissues" (Life Technologies, Cat no. 87790). The nuclear soluble proteins were obtained by following the instructions provided with the kit, except for changing the protease inhibitors included in the kit to the protease inhibitors (Roche, Cat no. 05892791001) and phosphatase inhibitors (Roche, Cat no. 04906845001). The nuclear soluble extract were supplemented with 5% glycerol (v/v), divided into 5 μl aliquots in each tube, snap froze in liquid nitrogen and stored at -80 °C for future usage.

3.2.2 Induced hepatocytes reprogramming assay

The 293FT cells were first transfected with lentiviral expression vector pLenti6/V5 with individual genes together with packaging vectors psPAX2 and pMD2.G (Addgene) using the Lipofectamine 2000 transfection reagents (Thermo Scientific, Cat no. 11668019) according to the manufacturer's instructions, and then replenished with fresh culture medium one day after the transfection. The lenti-viruses were harvested 48 hours after the transfection and further concentrated.

The early passage human fibroblasts were seeded on day 0, and then transduced with combinations of TFs on day 1 in the presence of polybrene (8 μg/ml final concentration; Sigma, Cat no. TR-1003-G). The medium containing viruses were changed to standard medium one day after transduction (day 2) and then changed to defined hepatocyte culture medium (HCM; Lonza, Cat no. CC-3198) on day 3. The medium was changed every other day, and on day 7 the cells were trypsinized and replated on type-I collagen coated plates in HCM medium. On day 29, the cells were again passaged to new type-I collagen coated plates and cultured until six weeks after transduction. The cocktails of TFs applied to the reprogramming assay were based on previous studies from Morris *et al.*[204] (FOXA1, HNF4A, KLF5), Du *et al*.[205] (HNF4A, HNF1A, HNF6, ATF5, PROX1, CEBPA), Huang *et al*.[206] (FOXA3, HNF4A, HNF1A), and the nine specific TFs identified by ATI from mouse liver (HNF1A, HNF1B, DBP, MAFG, CEBPA, CEBPB, HNF4A, HNF6/ONECUT1, ESRRA).

The cells were harvested at different time points, and the total RNA were extracted and followed by cDNA synthesis with the "Power SYBR™ Green Cells-to-CT™ Kit" (Thermo Scientific, Cat no. 4402954).

3.2.3 Active TF identification assay

The assay was performed *in vitro* by mixing 5 μl protein extract obtained from previous protein extraction step, 5 μl 140 bp double stranded DNA (dsDNA) oligos containing 40 bp random sequence in the middle (10 pmol) and adapter sequences on both ends, and 5 μl 3 × protein binding buffer (420 mM KCl, 15 mM NaCl, 3 mM K_2HPO_4, 6 mM $MgSO_4$, 300 μM EGTA and 9 μM $ZnSO_4$, 60 mM HEPES, pH = 7.5) and incubating for 30 min at room temperature. The non-specific competitor poly-dIdC was supplemented in the reaction (5 ng/ μl final concentration). After the reaction, the electrophoretic mobility shift assay (EMSA) was carried out using 6% DNA Retardation Gel (Invitrogen, Cat no. EC63652BOX) in 0.5 × TBE buffer (1 mM EDTA in 45 mM Tris-borate, pH 8.0) with 106 V constant voltage for 70 min. The gel above the 300 bp DNA marker was collected, eluted in 300 μl Tris buffer (10 mM Tris-Cl, pH 8.0) and incubated at 65 °C for 3 hours. The eluted DNA was amplified with Phusion polymerases (Thermo Scientific, Cat no. F530L); 4 pmol of each primer were used for the amplification. Before the final step of amplification, the same amount of primers was added to convert the single stranded DNA (ssDNA) to dsDNA. The amplified

DNA library was incubated again with aliquot of the same protein extract as above and the whole process was repeated for three more times. The PCR products with different barcodes were pooled and purified with QIAquick PCR Purification Kit (Qiagen, Cat no. 28106) for library preparation for next generation sequencing with Illumina instruments.

3.2.4 Bioinformatical analysis of ATI data

Two independent methods were utilized to analyze the ATI data to measure the enrichment of motifs, which reflected the total binding activities of corresponding TFs or TF families in the cell extract. The *de novo* motif discovery method utilized the "Autoseed" program[13] to detect the most significant motifs. Up to 200 highest count local maxima 10 bp sequences (with or without a gap at the center) were used as seeds to generate initial position weight matrix (PWM) motifs, which were further investigated manually to remove low complexity motifs and motifs that were highly similar.

The known motif enrichment analysis measured the enrichment of known motifs from the SELEX database[14,73] (including DNA-dependent dimeric motifs) based on the MOODS program[207,208]. First the frequency of each motif was calculated utilizing the MOODS program in cycle 1 and cycle 4 sequencing data with one cutoff (p-value ≤ 0.0001, Score > 11) based on the PWM of the motif. Subsequently, the enrichment and p-value (Winflat[209]) were calculated for each motif by comparing the frequencies of the motif in cycle 4 and cycle 1 sequencing data; the sensitivity to detect differences using this method was very high, and it could detect highly statistically significant differences whose fold-changes were probably too low to be detected by the *de novo* motif discovery method.

3.2.5 Analysis of DNase I hypersensitive sites (DHSs)

The DHSs data from different mouse tissues and ES cells were obtained from the ENCODE project[210], including 14 replicates for mouse liver, 7 replicates for mouse brain, 2 replicates for mouse heart, spleen and ES cells. The top 5,000 regions for each replicate were selected from the BroadPeak data set based on Signal Values. For samples with two replicates, the intersected regions were used for downstream analysis, resulting in around 4,000 DHSs for each tissue; for liver, DHSs overlapped by more than 8 replicates were selected to reach 3,806 DHSs, which were comparable with the other tissues; for brain, DHSs overlapped by more than 4 replicates were selected, resulting in 3,645 DHSs. Meanwhile, frequencies of all different 10-mers on both strands were counted in the ATI data; the fold change for each 10-mer was calculated by comparing the frequencies of it in the last cycle (Cycle 4) and the first cycle (Cycle 1). After that, the DHSs and the 10-mer results from the same tissues

were analyzed. First, each DHS site was flanked by adjacent genomic sequences (non-DHS regions) to achieve a 10 kb region, resulting in ~ 4,000 regions with the length of 10 kb for each tissue and cell type, all these 10 kb regions were then aligned by using the middle of the DHS as the center position. Within the 10-kb sequences, each position was regarded as a 10 bp sequence with that particular position as the 4th nucleotide; the score for each position was then calculated based on the log2 fold change of the corresponding 10-mer in ATI data.

The DHS sites were predicted from above mentioned 10 kb regions as well as from the whole genome. First, the scoring of the 10-mers was optimized by trying different cutoffs using a separate training set (setting separately top 0.1%, top 0.5%, top 1%, top 5%, top 10%, top 20%, top 40%, top 60% and top 100% of the 10-mers as score 1 and the remaining 10-mers as score 0, 100% of 10-mers is considered as a negative control), resulting in the optimal cutoff by setting the top 1% enriched 10-mers as score 1 and the remaining 10-mers as score 0. The 10 kb regions were divided to 50 bp bins, and each bin was then assigned with the mean score of all sliding 10-mers inside the bin. The DHS position was then called based on identification of the highest score in a sliding window of 17 bins. The optimal width of the smoothing window was determined by using half of the 10 kb regions as a training set; only the test set data is shown on **Figure 4.5c** for ES cells.

3.2.6 Capturing DNA-binding proteins using biotinylated ATI ligands

The DNA-binding proteins from the nuclear extract were captured as previously reported[211]. First, DNA oligonucleotides were amplified with biotinylated primers (modified with Biotin-TEG) and purified with QIAquick PCR Purification Kit (Qiagen, Cat no. 28106) to get rid of extra primers. Subsequently, 2 µg of biotinylated DNA library were incubated with 4 µl of high-performance streptavidin Sepharose (GE Healthcare, 17511301) in 100 µl DNA binding buffer (10 mM HEPES, pH 8.0, 1 M NaCl, 10 mM EDTA, and 0.05% NP40) for 1 hour at room temperature by shaking. Beads were then washed twice with 100 µl DNA binding buffer and twice with 100 µl protein binding buffer (140 mM KCl, 5 mM NaCl, 1 mM K_2HPO_4, 2 mM $MgSO_4$, 100 µM EGTA and 3 µM $ZnSO_4$, in 20 mM HEPES, pH = 7.5). 200 µl Nuclear extract from feeder-free mouse ES cells containing 200 µg proteins was mixed with the washed beads and incubated for 1.5 hours by shaking at room temperature. 2 µg poly-dIdC competitor DNA and EDTA-free complete protease inhibitors (Sigma, Cat no. 000000004693159001) were supplemented to the reaction. The beads were then washed with ice-cold low stringency buffer (10 mM Tris-Cl, pH 7.5, 4% glycerol, 500 µM EDTA, 50 mM NaCl) for 10 times and proceeded to on-beads digestion for MS[211].

3.2.7 Sample preparation for mass spectrometry

After capturing DNA-binding proteins from the nuclear extract, the washed beads were incubated in 50 µl of buffer containing 25 mM ammonium bicarbonate and 1 mM dithiothreitol (DTT) for 1 h at 37 °C, and followed by addition of 5 mM Iodoacetic acid (IAA) in the buffer for incubation in the dark for another 10 min. Extra DTT were supplemented to the final concentration of 5 mM to quench the reaction. The proteins were digested by directly incubating the beads with Lys-C protease (0.2 µg/sample; Thermo Scientific, cat. no. 90051) overnight at 37 °C, followed by digestion with trypsin protease (0.1 µg/sample; Thermo Scientific, cat. no. 90057) for another 10 hours at 37 °C.

Around 500 µg labeled peptide pool was dissolved in 250 µl of rehydration buffer containing 8 M urea and 1% Pharmalyte (pH 3–10, from GE Healthcare), followed by re-swelling an immobilized pH gradient (IPG) gel strip (GE Healthcare) at pH 3–10. The isoelectric focusing (IEF) was performed on an Ettan IPGphor isoelectric focusing system (GE Healthcare) to at least 150 kVh for around one day. After focusing, the MilliQ water was supplemented with the liquid-handling robotics (GE Healthcare prototype) for incubation/extraction of peptides. The extracted peptides were transferred into a microtiter plate (96 wells, V-bottom, Corning cat. no. 3894). The extraction was repeated three more times, and the combined samples on the microtiter plate were dried using a vacuum microcentrifuge.

3.2.8 Label-free mass spectrometry

Label-free mass spectrometry (MS) of peptides was operated using a hybrid Q-Exactive mass spectrometer (Thermo Scientific). The sample was resuspended in 10 µl of solvent A (5% DMSO and 0.1% formic acid (FA) in water), and 3 µl of it was injected. Peptides were trapped on an Acclaim PepMap nanotrap column (C18, 3 µm, 100 Å, 75 µm × 20 mm) and separated by an Acclaim PepMap RSLC column (C18, 2 µm, 100 Å, 75 µm × 50 cm, Thermo Scientific). Separation of peptides were performed by using a gradient of solvent A and solvent B (90% acetonitrile (ACN), 5% DMSO and 0.1% FA in water), with solvent B ranging from 6% to 37% in 240 min with a flow of 0.25 µl/min. Q-Exactive (QE) was performed in a data-dependent manner. First the FTMS (Fourier Transform Mass Spectrometry) survey was performed by scanning at 70,000 resolution (and mass range 300–1,700 m/z) followed by MS/MS (35,000 resolution) of the top five ions using higher-energy collision dissociation (HCD) at 30% normalized collision energy. Then precursors were isolated with a 2-m/z window, by setting the automatic gain control (AGC) as 1×10^6 for MS1 and 1×10^5 for MS2. Maximum injection times were 100 ms for MS1 and 150 ms for MS2. The entire duty cycle lasted around 1 s, and dynamic exclusion was applied with a 60s duration.

Precursors with unassigned charge state or a charge state of 1 were excluded, and underfill ratio of was set as 1%.

3.2.9 Peptide and protein identification

To determine the unique peptides and proteins, MS raw files were searched with the Sequest-percolator in Proteome Discoverer 1.4 software (Thermo Scientific) against the Uniprot mouse database (version 2016_10, canonical and isoforms, 85,832 protein entries) and filtered to a 1% false discovery rate (FDR) cut-off (peptide-spectrum-match level). A maximum of two missed cleavages was used together with: carbamidomethylation (C) set as fixed modification and oxidation (M) set as a variable modification. A precursor ion mass tolerance of 10 p.p.m. and a product ion mass tolerance of 0.02 Da for HCD spectra were applied. The average area of the top three peptide spectrum matches (PSMs) for each protein group was used to calculate protein area. Only unique peptides in the data set were used for quantification.

3.2.10 Reconstitution of nucleosomes

The nucleosomes were reconstituted as described previously[212]. First, 100 ng dsDNA were incubated with 50 ng histone octamers tagged with streptavidin binding peptide (SBP) in 2 M KCl solution (10 µl) for 30 min at room temperature. The solution was diluted stepwise by addition of dilution buffer (TE buffer supplemented with 1 mM tris (2-carboxyethyl) phosphine (TCEP) and a cocktail of protease inhibitors) until the final KCl concentration reached around 140 mM. The volumes of the subsequent buffer additions were 10 µl, 5 µl, 5 µl, 5 µl, 5 µl, 60 µl, 40 µl, each followed by one hour incubation. The DNA oligos were designed to contain 101 bp random sequences in the middle and flanked by 24 bp and 22 bp adapter sequences on both ends, resulting in the total length of 147 bp.

3.2.11 ATI assay by using nucleosomes

The reconstituted nucleosomes were first immobilized by the streptavidin-coated magnetic beads (28-9857-99, GE Healthcare; pre-blocked with the blocking buffer containing 25mM Tris, 0.5% BSA, 0.1% tween 20, 0.02% NaN_3) and washed using the protein binding buffer which was used in the standard ATI assay. Nuclear proteins extracted from particular cell types were added directly to the beads and incubated at room temperature for 90 min. For one single reaction, the total volume was 20 µl, including 0.5 pmol nucleosomes, 20 µg nuclear extract, and 5% w/v PEG4000 (Sigma, Cat no. 1546569-1G) in the same protein binding buffer used in the standard ATI assay. The DNA ligands dissociated from the histones were collected

from the supernatant. Meanwhile, the complex still bound to the beads were eluted by protein binding buffer supplemented with 10 mM biotin, and subjected to EMSA to separate the ligands that were bound by both histones and proteins from the nuclear extract. The ligands both from the supernatant and separated by EMSA were separately PCR-amplified and used for next cycle; the process was repeated for two more times.

4 RESULTS

4.1 STUDY I: DECIPHERING MOST ACTIVE TFs BY ATI

4.1.1 Extraction of nuclear soluble proteins

Since the binding activities of TFs are important for the cell identity, it is necessary to quantify the activities of different TFs located in the cell nucleus in order to decipher the mechanism of cell fate determination. Due to the development of technologies to study the binding specificity of TFs, especially the high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX), we have determined the binding specificity of most TFs in humans[71] and flies[13]. In this project, we established a novel method, active TF identification (ATI) assay to measure the DNA-binding activities of all TFs in cell nucleus.

Firstly, the nuclear soluble proteins were extracted from the cultured cells or fresh tissues. The proteins located outside the nucleus are not responsible for the transcriptional regulation occurring within the nucleus and therefore the cytosolic fraction was discarded. In addition, the extraction of nuclear soluble proteins was carried out at 400 mM salt concentration in order to extract most TFs together with epigenetic regulators but not the histones (**Figure 4.1**). Histones possess a certain degree of specificity in their DNA binding and as such could generate considerable noise in the assay.

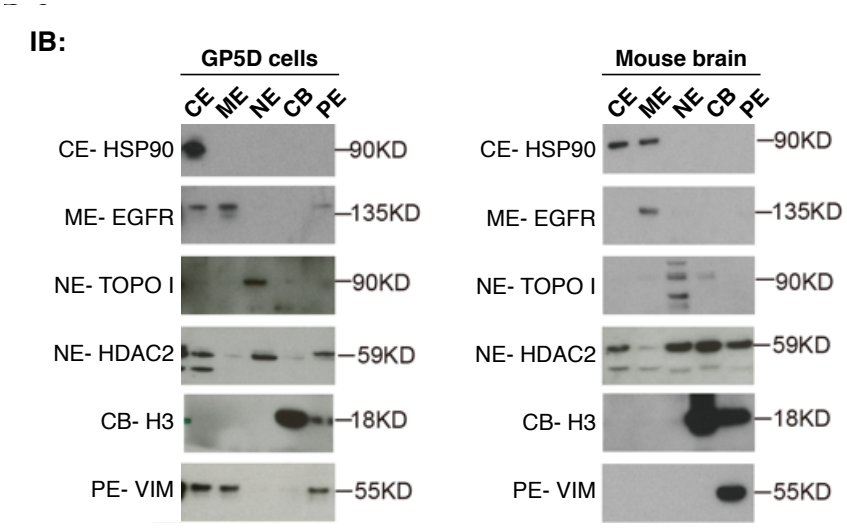

Figure 4.1 Subcellular protein extraction from cultured cell lines and fresh tissue

Immunoblotting assay (IB) was performed on extracts from GP5D human colorectal cancer cells and mouse brain tissue. The antibodies used are against proteins from the five different compartments including HSP90 from cytoplasmic extract (CE), EGFR from membrane extract (ME), TOPO I and HDAC2 from nuclear soluble extract (NE), histone H3 from chromatin-bound extract (CB), and VIMENTIN (VIM) from cytoskeleton extract (PE).

4.1.2 Active TF identification (ATI) assay

Next, the ATI assay was performed by incubating a library of double-stranded DNA containing 40 bp random sequences with a nuclear extract from particular cell or tissue type. The DNA bound by TFs was then separated from the unbound DNA by electrophoretic mobility shift assay (EMSA, **Figure 4.2a**). The bound DNA ligands were eluted from the gel and amplified by PCR, and the entire process was repeated three more times. Comparison of millions of sequences derived from the input and the selected libraries then allows identification of enriched binding motifs that correspond to the TFs present in the nuclear extract. The logic of this method is similar with proteomics, as each binding motif is regarded as the fingerprint of the particular TF or TF family, and the total binding activity for each individual TF or TF family can be determined by the total enrichment of the corresponding motif after four cycles selection in ATI. One concern is that the binding motif may be corresponding to one single TF or a group of TFs sharing the same DNA binding domain (DBD). In order to determine which TF from the same group is responsible for the enrichment of the related motif, we supplemented the ATI assay with a mass spectrometry identification process after capturing proteins from the same cell types by using the enriched DNA library (**Figure 4.2a**).

We applied the ATI method to different cell or tissue types, including mouse ES cells, heart, spleen, brain and liver dissected from one year old wild-type male mouse. By using the Autoseed program for *de novo* motif mining, we are able to detect around 10 distinct motifs for each sample (**Figure 4.2b**). Among these motifs, five of them are commonly active in all the cell or tissue samples tested, indicating that the corresponding TFs/TF families, including NRF1, NFI, bHLH, Nuclear Receptors (NR) and bZIP, are generally active in all cell types, acting as the housekeeping TFs. Three motifs are detected in more than two out of five samples, corresponding to TFs/TF families YY1/2, ETS and RBPJ; identification of these "shared" TFs/TF families indicates that the transcriptional regulatory network cannot be totally unique for particular cell identity, and part of the network is shared by other cell types. Moreover, we also detected "specific motifs" in only one or two out of five samples, indicating the unique roles of the corresponding TFs/ TF families in the specific cell identity. In addition, we also detected enrichment of some unknown motifs (**Figure 4.2b**, bottom), which we could not assign to a known TF based on the current knowledge (HT-SELEX motifs, CIS-BP, TOMTOM[12,13,71,213]). Overall, we recovered 35 motifs, of which only six (17%) were unknown, indicating that specificities for most of the TFs that display strong activity in the tested tissue types have already been determined.

In addition to discovery of TFs binding motifs, the quantitative information on binding activity of TFs can also be obtained from the *de novo* motif mining method. By comparing the enrichment of the "common" motifs shared by all the samples tested (**Figure 4.2c**), we found that the binding activity of the same TF/TF family varied a lot in different cell or tissue types, implying additional specific roles of them in different

cell types. For instance, the NFI family motif was most highly enriched in the liver, indicating that the NFI family TFs in liver not only perform fundamental duties as the house keeping TFs, but also regulate expression of liver specific genes as the liver specific TFs; in addition, the enrichment of NRF1 motif was extremely high in spleen compared with other cell and tissue samples, implying the specific role of TF NRF1 in spleen. In general, our ATI results provide both qualitative and quantitative information about the binding activity of TFs.

Apart from the *de novo* motif mining, we also performed known motif enrichment analysis on the ATI sequencing data by means of MOODS program, which calculated the enrichment of all known motifs from our SELEX database in the enriched DNA library (cycle 4) compared with the DNA library after one round selection (cycle 1). The enrichment of the motif reflects the binding activity of the corresponding TF/TF family in cells. In general, known motif enrichment analysis results were consistent with motifs discovered by using the *de novo* discovery method, and revealed additional TFs whose motifs were specifically enriched in different cell types or tissues. For example, motifs for pluripotency factors such as GLIS2 were detected in mouse ES cells with the known motif enrichment analysis, albeit with a relatively low enrichment; from the mouse liver, we also detected motifs for liver specific TFs such as ONECUT and HNF4A (**Figure 4.2b**).

Taken together, the ATI results from the tested cell and tissue samples revealed that in addition to five commonly active TFs, each tissue or cell type displayed strong DNA-binding activity of key regulators of the respective cell identity.

4.1.3 Identifying specific transcription factors by MS

The ATI technique is quite useful to identify the most active motifs in different cell types, but in most cases, it cannot identify the specific TFs that are active in cells, due to the fact that many TFs within the same structural family share the same binding motif. In order to identify the specific TFs that cannot be determined by their binding motifs only, we captured the proteins from nuclear extract of mouse ES cells by using the control and enriched ATI DNA libraries, and performed mass spectrometry (MS) analysis on the captured proteins. Based on the MS results, we managed to identify specific TFs corresponding to active motifs (**Table 4.1**). For example, the POU family protein POU5F1 was detected with high abundance in the ES cells, indicating that POU5F1, rather than other POU family proteins was predominantly active in the ES cells; the MS result also revealed that the KLF4 protein was the most active KLF/SP family TF in the ES cells. In most cases, the abundance of TFs captured by the enriched DNA library was higher than those enriched by the original DNA library, indicating that the DNA ligands containing specific motifs from the original library are fully captured by corresponding TFs, but those ligands from the enriched library may not be saturated. However, there was one exception- the abundance estimate of ZIC family

TFs was higher when using the original DNA library than using the enriched DNA library, which might be caused by random variation of the MS assay.

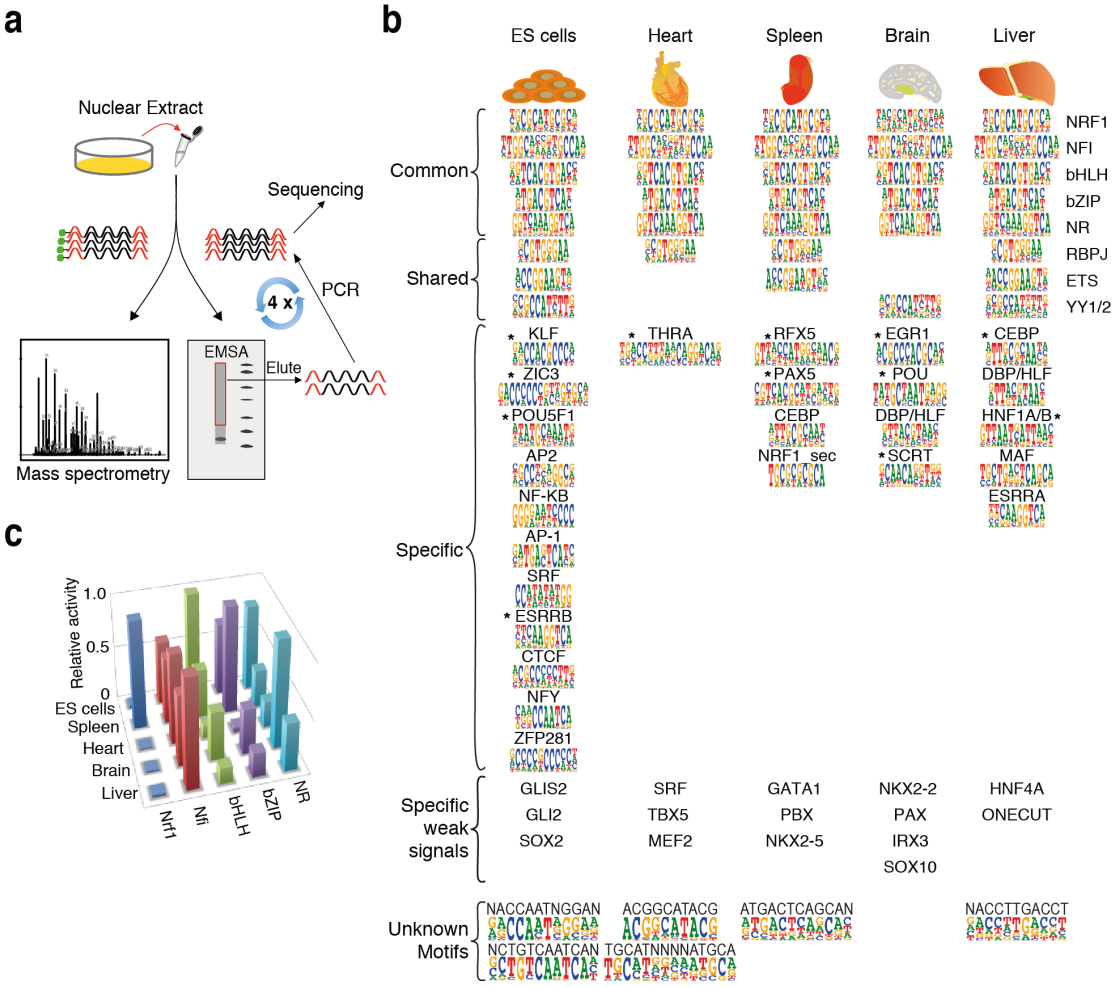

Figure 4.2 Determining the most active TFs in cells by means of ATI assay

a) Schematic description of the active transcription factor identification (ATI) assay. A pool of dsDNA ligands is incubated with nuclear extract of cells or tissues; the oligos bound by the nuclear proteins are selected against the unbound ones by native PAGE gel purification (EMSA; right) and amplified. The process is done in four cycles, resulting in an enriched DNA pool that reflects binding activities of the TFs in the cell lysate. Both the original and enriched DNA libraries are subjected to next generation sequencing for motif analysis. TFs are also captured by incubating the nuclear extract and a biotinylated oligonucleotide pool consisting of the identified ligands, and followed by MS (left) for quantification of the proteins.

b) The most highly enriched motifs and the corresponding TFs identified by *de novo* motif discovery from different cell or tissue types. The TFs known to contribute to lineage determination in the analyzed samples are indicated by asterisks. The names of the TFs are based on the motifs. In cases where multiple TFs share the same binding motif, the representative TF is indicated based on the mRNA expression levels and functional data from previous studies. Examples of TFs known to be important for the specific tissues whose motifs were identified by only using the known motif discovery pipeline are also indicated in "specific weak signals". Some detected but unknown motifs are also shown on the bottom.

c) Variation of DNA-binding activities for common TFs in different tissue or cell types. The bars represent the relative binding activities of the "common" motifs in different tissue or cell types, which is based on increase of absolute molecular counts[214] of each motif between the original library (cycle 0) and the selected library after the last cycle (cycle 4). The activities of each TF were normalized by setting its highest activity in any of the tissues to 1.

Table 4.1 Identification of specific TFs based on the MS result

TF detected in ATI	TF detected in MS	Protein area in Cycle 0 (C0)	Protein area in Cycle 4 (C4)	Fold change (C4/C0)
RBPJ	RBPJ	2,81E+09	3,86E+09	1,37
bHLH	USF1	7,94E+08	2,14E+09	2,70
NRF1	NRF1	2,74E+07	3,24E+07	1,18
YY1/2	YY1	3,04E+08	3,55E+08	1,17
AP2	TFAP2C	4,49E+08	4,71E+08	1,05
CEBP	CEBPG	9,12E+07	2,09E+08	2,29
KLF/SP	KLF4	3,19E+08	4,21E+08	1,32
POU	POU5F1	5,59E+08	7,13E+08	1,28
NFY	NFYA	3,12E+08	4,85E+08	1,55
NFI	NFIB	6,96E+06	2,25E+07	3,24
NR	RARG	1,56E+08	1,67E+08	1,07
ETS	ERF	1,30E+08	1,36E+08	1,05
ZIC	ZIC3	6,64E+08	3,94E+08	0,59
ESRR	ESRRB	2,32E+09	2,22E+09	0,96
RFX	RFX1	1,65E+08	2,80E+08	1,69
SRF	ND	N/A	N/A	N/A

ND: Not determined

4.1.4 Binding activity changes during differentiation

To study the activity changes of TFs during dynamic processes, we applied the ATI assay to compare the TF binding activity in ES cells and the differentiated cells (**Figure 4.3a**). The mouse ES cells were induced towards neural and mesodermal lineages using standard conditions[215-217], and the enrichment of each motif in different lineages was compared with its enrichment in ES cells, revealing the binding activity changes during the differentiation. Because the original DNA libraries are the same for different samples, relative enrichment for each motif in different samples is then equal to the ratio of the frequencies for each motif in corresponding cycle 4 DNA libraries. The results indicated that several known quantitative changes in TF binding activities accompanied the neural and mesodermal differentiation processes. For instance, the activities of the pluripotency factors GLIS and ZIC were decreased, whereas the activities of RFX and PAX factors that are known to contribute to neural differentiation were increased (**Figure 4.3b** and **c**). Similarly, the activities of GLIS and ZIC factors decreased after induction of mesodermal differentiation, whereas the activity of the known mesodermal factor AP2 increased significantly (**Figure 4.3b** and **d**). However, the activation of SMAD proteins by BMP4 and Activin A that were used to induce mesodermal differentiation was not detected, potentially due to the fact that interactions between SMAD proteins and the DNA are too weak ($K_d \approx 1 \times 10^{-7}$ M), and often rely on some other TFs[218]. In contrast, ATI robustly detected the activation of retinoic acid

receptor, one type of nuclear receptor (NR), by the neural inducer retinoic acid (**Figure 4.3b** and **c**), indicating that some ligand-inducible TFs can be detected by ATI.

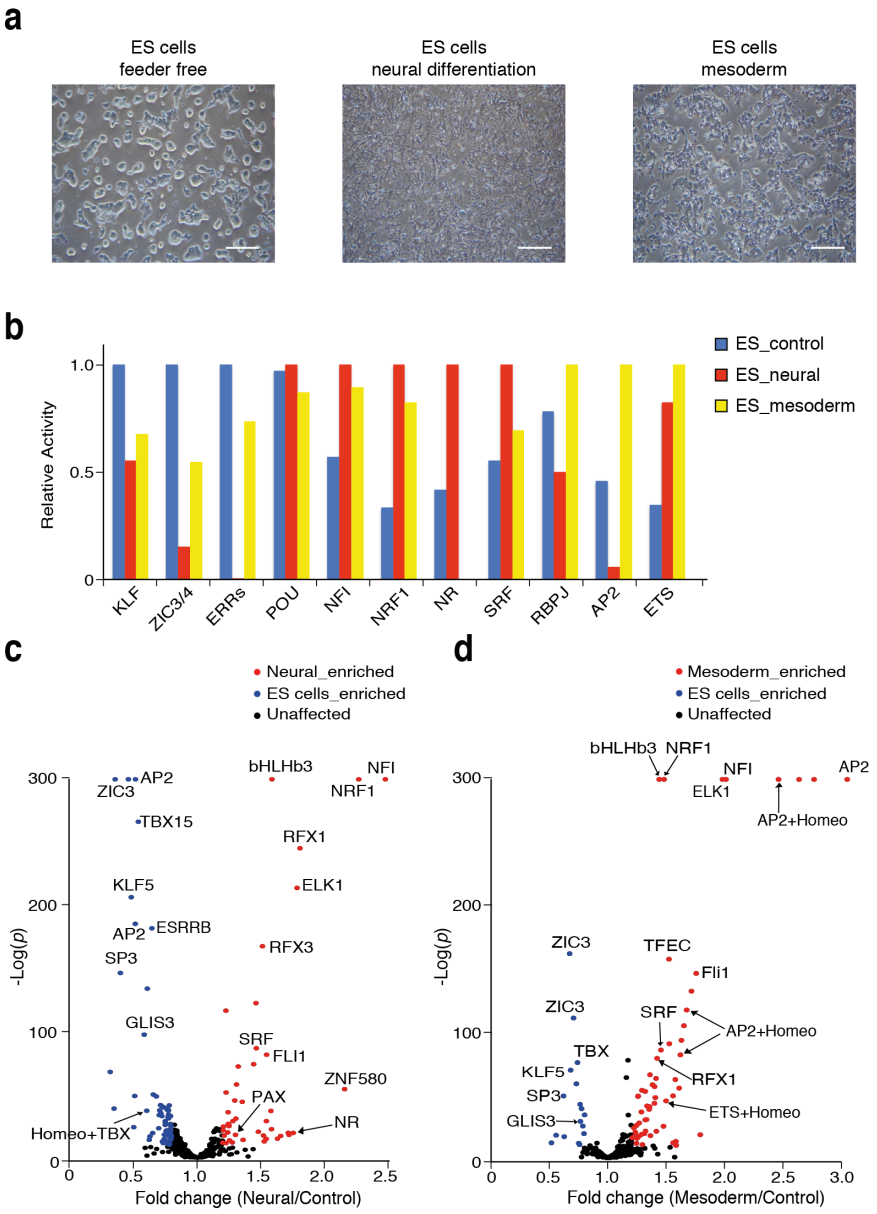

Figure 4.3 Analysis of TF activity changes during ES cell differentiation

a) Morphology changes of mouse ES cells (left) differentiated into neural (middle) and mesodermal (right) lineages. Scale bars, 400 μm.

b) Comparison of binding activity of TFs /TF families detected by the *de novo* motif discovery method in the control and differentiated mES cells. Bars indicate the relative activities of the indicated TFs/TF families based on increase of the absolute molecular counts[214] of the corresponding motif from cycle 1 to cycle 4. For each TF/ TF family in different lineages, their activities were normalized by setting the highest activity in any of the three conditions to 1.

c, d) Comparison of motif enrichment between the neural (**c**) or mesodermal (**d**) differentiated ES cells and the control undifferentiated ES cells by using the known motif frequency analysis method. The y axis indicates the *p* value (log scale, calculated by winflat[219]; owing to the precision of calculation, many *p* values were set to a minimum of 1×10^{-300}); the x axis indicates fold change. Motifs with $p < 1 \times 10^{-10}$ and more than 20% changes are indicated in red (enriched in neural or mesodermal differentiated ES cells) or blue (enriched in control ES cells), respectively, with the names of representative motifs indicated. The black dots represent motifs that changed less than 20% and/or did not pass the *p*-value threshold.

4.1.5 Reprogramming of induced hepatocytes with overexpression of specific TFs detected in ATI

In order to prove that the TFs we have found in the ATI assay are the master regulators of the specific cell identities, we performed the reprogramming assay to convert the fibroblasts to hepatocytes by overexpressing nine TFs that were specifically detected in adult mouse liver (Set_ATI), namely HNF1A, HNF1B, CEBPA, CEBPB, DBP, MAFG, ESRRA, HNF4A and HNF6/ONECUT1, and investigated the morphology of the cells and expression of the liver specific marker gene *ALBUMIN* after two weeks of culture. We also performed the reprogramming assay using other combinations of TFs that were previously published[204-206] (Set_a, Set_b and Set_c).

The result indicated that by overexpressing those nine liver specific TFs, the fibroblasts were successfully converted to induced hepatocytes (iHeps) two weeks after transduction, in a manner comparable to the other three methods described previously. Moreover, the expression level of *ALBUMIN* in Set_ATI was one of the highest among all four sets, indicating its high efficacy. Taken together, our iHeps reprogramming assay revealed that by introducing nine liver specific TFs detected in the ATI assay, we were able to convert the fibroblasts to hepatocytes at an efficiency similar to that of the most efficient previously described protocol.

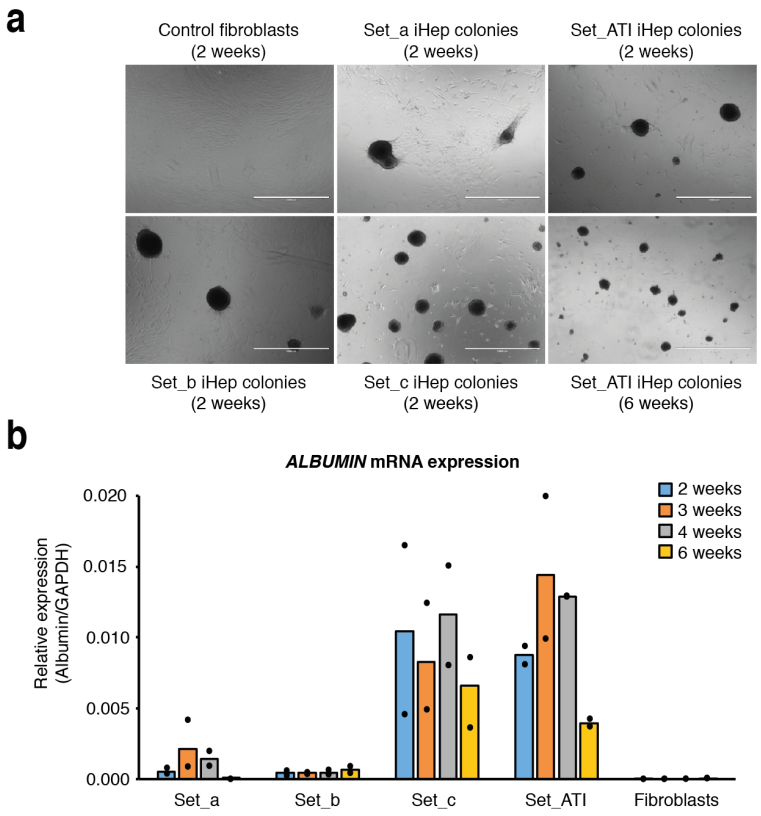

Figure 4.4 Reprogramming of human fibroblast to induced hepatocytes

a) Bright field images of iHep colonies from human fibroblasts after lentiviral transduction of TF combinations previously reported in Morris *et al.*[204] (Set_a; FOXA1, HNF4A, KLF5), Du *et al.*[205] (Set_b; HNF4A, HNF1A, HNF6/ONECUT1, ATF5, PROX1, CEBPA), Huang *et al.*[206] (Set_c; FOXA3, HNF4A, HNF1A) and factors identified by ATI in mouse liver (Set_ATI; HNF1A, HNF1B, DBP, MAFG, CEBPA, CEBPB, HNF4A, HNF6, ESRRA). Scale bars, 400 μm.

b) Relative expression levels of the liver-specific marker gene *ALBUMIN* in iHep cells normalized to *GAPDH* levels by RT-PCR using previously reported TF cocktails and ATI-identified TF combinations. Bars indicate the means of two independent duplicate samples, and the dots indicate values for each duplicate sample.

4.2 STUDY II: CORRELATION BETWEEN DHSs AND ATI DATA

Due to their binding specificity with DNA, TFs are considered to be the most important *trans*-acting factors required to set up the chromatin landscape in cells. In order to determine whether ATI also confers information on the mammalian chromatin landscape, we compared the ATI data with DNase I hypersensitive sites (DHSs) in mouse ES cells from the mouse ENCODE project[210]. This analysis revealed that the top 2000 most enriched 10-mers detected by ATI in mouse ES cells were strongly enriched in the ~ 5000 most significant DHS regions from the ES cells (**Figure 4.5a**). Moreover, we also performed the same analysis on other mouse tissues, resulting in significant but weaker enrichment than the ES cells (data not shown). The reason may be that ES cells are more homogenous than tissues containing multiple different cell types. Further analysis of 10-mers enriched in both DHSs and ATI data from ES cells revealed that there were many 10-mers that were enriched in both, and that all of these 10-mers were related to the ATI motifs (**Figure 4.5b**). However, there were some other 10-mers only enriched in DHSs (**Figure 4.5b**), including many repetitive CG rich sequences that were enriched in gene regulatory elements due to the fact that methylated C is prone to mutation, and the low CpG methylation rate of regulatory elements protects these sequences from this mutational process[220,221].

We further hypothesized that we could predict the DHS regions using the ATI enriched subsequences since ATI could accurately represent TF binding activities in cells and revealed subsequences that bound strongly to TFs *in vivo*. It has been well accepted that DHSs and TFs binding clusters are enriched with matches to biochemically obtained TF motifs, and that they overlap with *in silico* predicted clusters called based on TF motif matches only[222,223]. However, studies in our lab have shown that only ~ 30% of TF binding clusters could be predicted based on monomeric TF binding models[222], suggesting that other unknown determinants played more important roles in TF binding to DNA *in vivo*. In order to test if ATI can improve the prediction, we developed a predictor based on the enrichment rank of all 10-mers in ATI data. The results revealed that more than 70% of the DHSs could be predicted by the 10-mers derived solely from ATI data (**Figure 4.5c**; 10% expected by random, $p < 3.2 \times 10\text{-}226$; winflat[209]), indicating that ATI derived 10-mer enrichment more accurately represented TF activity as well as chromatin accessibility in cells compared to any other presently available information. In the analysis of genome-wide DHSs prediction, the ATI data was nearly as effective as using 10-mers from the DHSs themselves (**Figure 4.5d**), indicating that the it contains substantial fraction of the motif information included in the DHSs, despite the fact that the DHSs are expected to contain additional motif features that relate to their functionality in gene regulation and not to their open chromatin status. For example, *de novo* motif discovery analysis of the DHSs in ES cells detected many motifs that were different from the ones detected in ATI, including one motif similar to that of Znf-143 (**Figure 4.5e**); this motif has been reported to contribute to interactions between promoters and distal regulatory elements[224].

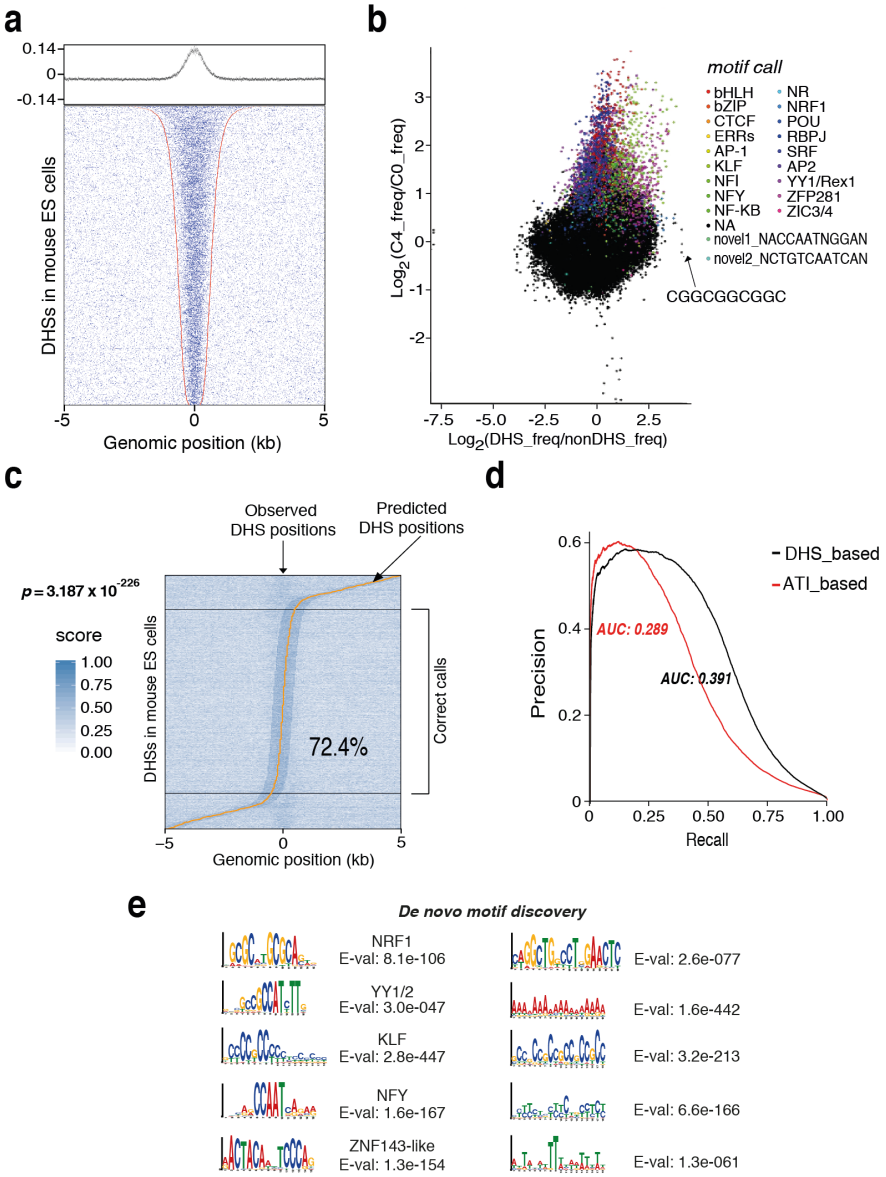

Figure 4.5 See next page for caption

Figure 4.5 Correlation of the DHS data and ATI results from mouse ES cells

a) The ATI-enriched 10-mers in ES cells are also enriched in DHSs from ES cells. In the dot plot, the rows indicate 3,907 most significant DHS sites and flanked genomic sequences from in mouse ES cells; the length of each DHS site together with flanking regions is 10 kb. In each row, red dots indicate the two boundaries of the DHS region; blue dots indicate the positions of the top 2% ATI-enriched 10-mers. The graph on top shows the average of scores for each 10-mer at each position across the rows.

b) Comparison of 10-mers in ATI data and DHS regions from mouse ES cells. X-axis indicates the log2 fold change of 10-mer counts in DHSs compared with non-DHS regions; y-axis indicates log2 fold change of 10-mer counts in ATI enriched DNA pool (Cycle 4) compared with original pool (Cycle 0). Colored dots represent 10-mers that are similar to the motifs detected in the ATI assay; black dots indicate the 10-mers that are not similar to any motifs. One 10-mer sequence ("CGGCGGCGGC") is shown as an example of repetitive CG rich sequences which displays high enrichment in DHSs but no enrichment in ATI.

c) Prediction of ES cells DHS regions based on the ATI result. DHSs were sorted by position of the prediction call (yellow line). Black horizontal lines separate accurate DHS calls (in the middle of the plot) from calls >500 bp off the known DHS center, which is located at the x-axis position 0 in all DHS sites. The fraction of predictions within ±500 bp of the center (72.4%), and the corresponding *p* value (Winflat) for the null model in which position calls are randomly distributed are indicated.

d) Genome-wide predictions of the ES cell DHS regions using 10-mer data from the ATI assay and the DHSs themselves are shown. The black line represents prediction based on the DHS data (details are included in the **"Meterials and Methods"** section), and the red line represent prediction based on the ATI data. AUC means the area under the curve, and indicates the accuracy and sensitivity of the method.

e) The top ten motifs detected with the lowest E values in DHS regions from ES cells are shown.

4.3 STUDY III: THE ATI ASSAY USING NUCLEOSOMAL DNA

4.3.1 Reconstitution of nucleosomes with DNA ligands

Although the ATI technology is quite powerful to determine the most active TFs for specific cell identity, it is not so efficient to detect the pioneer TFs with high binding affinity with nucleosomal DNA to initiate chromatin structural changes at specific loci. In order to determine the pioneer TFs in different types of cells, we modified our ATI assay by using nucleosomal DNA instead of naked DNA to incubate with the nuclear extract from different cell types.

The nucleosomes were reconstituted as described previously[212] using the histone octamers tagged with SBP and 147 bp dsDNA with 101 random bases in the middle. Subsequently the reconstitution mix was diluted to a salt concentration of 140 mM representing the physiological condition, a small aliquot was removed for EMSA experiment, and the remaining nucleosomes as well as the histones were immobilized by addition of streptavidin magnetic beads. The free DNA not bound by the histones was then washed away.

The EMSA result confirmed that well-organized nucleosomes were reconstituted that shifted the bound DNA fragments to around 400 bp DNA marker in the gel. The free DNA migrated at the expected 150-bp position (**Figure 4.6**). Quantitative analysis of the EMSA result further revealed that more than 50% of the DNA ligands were assembled into nucleosomes.

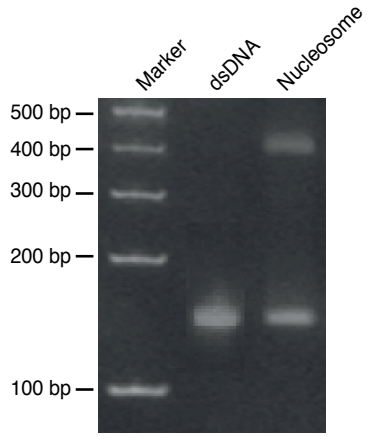

Figure 4.6 Reconstitution of nucleosomes with 147 bp dsDNA

The EMSA result shows the reconstituted nucleosomes. The free dsDNA is shifted at approx. 150 bp, and the reconstituted nucleosomes are shifted to 400-bp position.

4.3.2 ATI assay with nucleosomal DNA determines pioneer TFs

In order to determine the TFs that can compete with histones to bind DNA with high affinity, we incubated the reconstituted nucleosomes with the nuclear extract from mouse ES cells and liver tissue, and then collected the DNA disassembled from the nucleosomes (**Figure 4.7a**, "supernatant") and nucleosomal DNA still bound to the nuclear proteins by means of EMSA (**Figure 4.7a**, "EMSA shifted"). The DNA ligands collected were amplified and the whole process was repeated for two more cycles.

De novo motif analysis of both types of DNA libraries based on Autoseed program detected several motifs, some of which were not even detected in the standard ATI assay by using the pure DNA, implying that the TFs bound to these motifs play "pioneer" roles in determining the cell fate. For instance, the bHLH tetrameric and HOMEO-domain binding motifs were detected with high enrichment in ES cells and liver in both "supernatant" and "EMSA shifted" conditions, but they were not detected in the standard ATI assay. Moreover, in the ES cells, we also detected the binding motif for heterodimers formed by POU and SOX families TFs under both conditions. In addition, we found motifs that were also detected in the same tissue or cell samples in the standard ATI assay, for example, the bZIP and ETS motifs in ES cells, and the CEBP and NRF1 motifs in the liver.

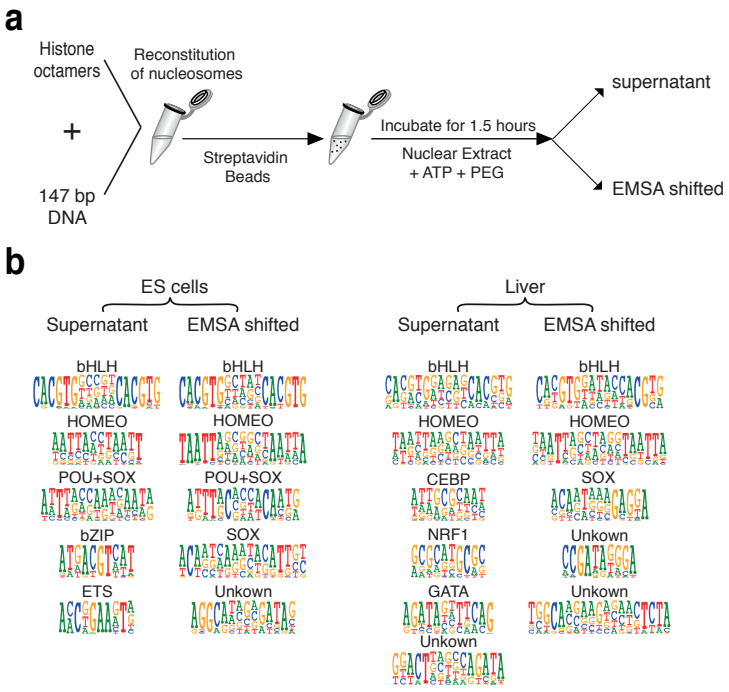

Figure 4.7 ATI assay using nucleosomal DNA determines pioneer TFs

a) Schematic of the ATI assay with nucleosomal DNA. ATP supplemented provides energy to dissociate the nucleosomes; PEG is used to concentrate the proteins and DNA by decreasing the volume of reaction.

b) Sequence logos and the corresponding TF families identified by *de novo* motif discovery from supernatant and EMSA shifted fractions from ES cells and liver tissue. The names of the TFs are based on the identified motifs. Motifs which cannot be determined based on the current database are marked as "Unknown".

5 DISCUSSIONS

5.1 LIMITED SETS OF TFs ARE HIGHLY ACTIVE IN SPEICIFIC CELL IDENTITY

Through measurement of the DNA-binding activities of all TFs in different cell or tissue types, the ATI assay indicates that limited sets of TFs are highly active in specific cell identity.

As one of the most important proteins involved in gene regulation, the TFs play dominant roles in determining specific cell identity. Nowadays, there exist a number of sophisticated technologies that can be applied to study various aspects of TFs. For example, as has been mentioned before, *in vitro* techniques such as the SELEX are applicable in determining the binding specificities of different TFs; moreover, the SELEX assay can also be utilized to study the kinetics of TF binding. The *in vivo* technique ChIP is applicable in detecting individual binding events of TFs as well as other types of proteins in the genome, which directly relates to transcriptional regulation controlling particular cell identity.

The strategy of the ATI technique is quite similar with that used in proteomics. The proteomics technology is applied to measure abundance of proteins based on quantification of unique peptides (the fingerprints) for all the proteins. Similarly, the ATI technique is applied to massively measure the total binding activities of different TFs or TF families in the cell nucleus based on the fingerprints- the specific binding motifs of different TFs or TF families.

Differing from techniques such as SELEX and PBM based assays which are utilized to determine TF binding specificity *in vitro*, the ATI assay measures the binding activity of TFs/TF families that exist in the cell nucleus based on the enrichment of their binding motifs. The word ‘activity’ is used here in the same sense as in enzymology, where activity represents total enzyme activity (specific activity × molar amount), thus the binding activity of TFs measured in the ATI assay is dependent not only on the abundance, but also the specific binding activity of the TFs. The results obtained from the ATI assay are unique and more useful than from other techniques such as RNA-seq or proteomics, because compared with expression levels, the binding activity of TFs is more directly related to the functionality of TFs. Additionally, the influence of post-translational modifications of TFs on their DNA binding activity cannot be assessed solely from their expression levels. With the ATI method, we are now finally able to determine the most active TFs in different cell types which was hitherto impossible to achieve with other high-throughput technologies such as RNA-seq.

In addition to massively measuring the binding activities of different TFs in the cell nucleus, the ATI assay is also quite useful to study the dynamic changes of TFs’ binding activities under perturbations. For instance, the ATI assay can been applied to

study the TF binding activity changes during differentiation from the ES cells to different lineages, resulting in the identification of lineage specific TFs.

The ATI result suggests that a small number of TFs with the highest DNA binding activities in a cell play a major role in setting its overall gene regulatory architecture. On the other hand, based on the ChIP-seq analyses, it has been proven that plenty of TFs actually bind open chromatin regions in the same cell identity[71,225]. These observations are consistent with a model where TFs that are strongly active in DNA binding set up the overall chromatin landscape of the genome, and the binding of TFs with weaker DNA binding activity is conditional on this chromatin landscape. This gene regulatory model is quite hierarchical, and is consistent with the hierarchical gene expression patterns commonly observed in analyses of real biological systems. In addition, this model also provides a simple combinatorial gene regulation system. If the TF that has strong DNA-binding activity lacks a strong transactivation or repression domain, it will require a partner that has such a domain. This cooperating factor may not, in turn, be able to bind DNA strongly to open chromatin alone, and therefore will require the strong DNA binder. It should be noted that different types of activation domains can also contribute to such combinatorial regulation, increasing the number of cooperation partners to three or more.

Admittedly, there are several weak points of the ATI technique. First, the ATI assay failed to detect the DNA-dependent cooperative binding of multiple TFs, which is frequently detected *in vivo*. The reason may be that in living cells the proteins in the nucleus are more concentrated than the proteins extracted from the cells; moreover, it has been stated that different kinds of biomolecules such as the transcriptional coactivators[226] and the RNA-binding proteins[227,228] are more concentrated at specific regions due to liquid-liquid phase separation of the proteins, implying the same effect for TFs. In addition, the ATI assay is not so sensitive to detect the pioneer TFs because for many pioneer TFs such as FOXA1, the expression level is not high enough under normal physiological conditions to be detected in ATI, otherwise lots of condensed chromatin regions will be open owing to overexpression of such pioneer TFs, leading to diseases such as cancer. It is plausible that during transition of cell lineages in normal physiological conditions, the pioneer TFs are temporarily expressed to open particular loci in the genome, after which they will be down-regulated.

5.2 CORRELATION BETWEEN BINDING ACTIVITY OF TFs AND CHROMATIN ACCESSIBILITY

Analyses of the ATI data and the DHS regions from the same cell or tissue types indicate that the binding activity of TFs and chromatin accessibility in the same cell identity are closely correlated, meaning that much of the nucleosome-competing activity in cells is due to the TFs present in such a high abundance relative to their K_d values in the nucleus that they can effectively and specifically compete against

nucleosome binding. Based on the ATI data, we are also able to predict positions of open chromatin in mouse ES cells far more accurately than what has previously been possible, indicating that the knowledge of TF DNA-binding activity levels is a major unknown factor that hindered previous computational predictions of regulatory elements. Furthermore, these results suggest that strongly active TFs have a major role in setting up the overall chromatin landscape of cells. However, our results cannot be interpreted to mean that open chromatin would result exclusively from the action of TFs with strong binding activity, or that those strongly bound TFs would be sufficient to open closed chromatin states characterized by presence of HP1, histone H1 or repressive chromatin modifications[229]. It is well known that some TFs can directly or indirectly recruit enzymes that remodel or modify nucleosomes to generate open chromatin and/or de-repress closed chromatin states[230,231]. It should also be noted that quite a few binding sites for less active TFs, or site(s) for cooperatively bound TFs can also be bound with sufficient energy to dissociate nucleosomes. Through these mechanisms a subset of genomic loci will become accessible, but they are unlikely to be the dominant way to open chromatin, as if that was the case, the *de novo* motif mining of the DHS regions would be able to detect the corresponding motifs for those cooperative or weak DNA binders.

Moreover, as DHSs represent gene regulatory elements, they are expected to be enriched with not only motifs that contribute to opening of the chromatin, but also sequences that are responsible for downstream activities such as transactivation or recruitment of RNA polymerase II. Consistently, *de novo* motif discovery analysis of DHSs revealed some motifs that were not enriched by ATI. These included a motif similar to that of Znf-143 (**Figure 4.5e**). On the other hand, because the nucleosomes possess weak DNA binding specificity, they also contribute to the accessibility of the chromatin genome-wide. However, the influence caused by nucleosome binding preference has nothing to do with the cell fate, as it is the same for all cell identities with the same genetic background.

In summary, the close correlation between ATI and DHS data verify that the TFs found using ATI are most active and important for the specific cell identity, and greatly contribute to genome accessibility in cells. The study of chromatin accessibility is still at the primary stage, as we mostly focus on the locations of the accessible regions and the sequence features of different DHS sites. Some effort should also be put on the classification and functionality of different DHS sites, and the correlation between each DHS site and epigenetic states at the adjacent regions.

5.3 PIONEER FACTORS WITH HIGH BINDING ACTIVITY WITH NUCLEOSOMES

By incubating nuclear extract from cells with nucleosomal DNA in the ATI assay, we detected TFs with the highest binding activity with nucleosomes, including some TFs that were not detected in the standard ATI assay. These detected TFs are regarded as "pioneer" TFs. It is noteworthy that the biochemical activity–based identification of pioneer TFs in this thesis is related, but not identical, to the classical concept of "pioneer" TFs, which is based on their functionality that they can access their target sequences in compacted chromatin and facilitate opening those regions[232]. Take the TF JDP2 as an example, it has been shown that JDP2 can bind the nucleosomes directly but promote the assembly of chromatin afterwards[133].

The pioneer TFs detected in the assay can be further divided into two groups based on their binding activity with the naked DNA. TFs including bZIP and ETS family TFs in ES cells, NRF1 and CEBP family TFs in the liver have significantly high binding activity with both the naked DNA and the nucleosomes; the motifs for these TFs are only detected in the "supernatant" (**Figure 4.7**), suggesting that their abilities to compete against the nucleosome are mainly due to the mass action rather than interacting with nucleosomes. Motifs for TFs such as those from SOX and HOMEO-BOX families can only be detected by using the nucleosomal DNA (**Figure 4.7b**), indicating that the corresponding TFs may have higher binding activity with the nucleosomes than the naked DNA. The reason why these TFs can bind nucleosomal DNA with higher affinity is not clear, partially due to the contact between the histone residues and the TFs, or the conformational change of the DNA.

In addition, it is obvious that motifs detected in the "supernatant" are related to dissociation of the nucleosomes, indicating the roles of their corresponding TFs in opening up condensed chromatin. However, motifs detected in the "EMSA shifted" component may be related to either condensation or opening of the chromatin, which is highly dependent on other recruited cofactors. In cases where motifs are detected in both "supernatant" and "EMSA shifted" components, for example the cooperative binding motif of POU and SOX family TFs detected in the ES cells, the contribution of their binding is also dependent on the cofactors they recruit at specific genomic loci.

The pioneer TFs are considered as the dominant factors to initiate changes of chromatin accessibility during cell fate transition, but none of their motifs were detected by the *de novo* motif discovery analysis of DHSs, indicating that the pioneer TFs are only responsible for opening a small subset of DHS regions in the genome. It is also possible that some pioneer TFs are activated temporarily for specific cell identity and then become inactive at a later time point, hence the binding activity of pioneer TFs at an earlier time cannot be reflected by the chromatin accessibility at a later time point. In order to identify those temporarily activated pioneer TFs, it is necessary to capture the specific states when the pioneer TFs are activated.

6 CONCLUSIONS AND PROSPECTS

All studies included in this thesis have greatly enhanced our understanding of how different types of TFs contribute to gene expression in specific cell types.

Although most TFs are expressed in each cell type, only a small subset of them plays a dominant role in determining specific cell fate. These dominant TFs, including both general TFs which are ubiquitously active in all cell types and specific TFs which are activated only in specific cell lineages, are responsible for setting up the core transcriptional regulatory network and regulate gene expression by interacting with other less active TFs and cofactors, leading to the hierarchical transcriptional regulatory network. With the development of the ATI technology, we are able to decipher the dominant TFs based on their DNA-binding activities. Moreover, because the TFs are so important for the cell fate, variation of their binding activities is highly correlated with transition of cell identities, which frequently occurs during many dynamic processes such as development and tumorigenesis. By means of the ATI assay, we are now able to massively measure the activity changes of all TFs in parallel and identify the crucial TFs with significant DNA-binding activity changes, which can hardly be achieved by other high-throughput technologies.

In addition, the thesis also demonstrates that the cooperative binding activity of TFs is a major determinant of chromatin accessibility for specific cell identity. Based on the ATI result, we can predict the positions of open chromatin in mouse ES cells more accurately than any method applied before, indicating that TFs, especially the strongest ones, play dominant roles in determining the chromatin landscape of cells in mammalian species. Admittedly, other factors such as the binding specificity of nucleosomes and the epigenetic modifications also contribute to the chromatin landscape. The nucleosome binding specificity determines for example the starting and ending points of transcription units to reduce inefficient transcription. The epigenetic modifications could affect the chromatin accessibility by recruiting chromatin remodelers or influencing TF binding at those sites. Moreover, it is stated in this thesis that the dominant TFs identified in specific cell identity through the ATI technology are regarded as the key factors which can be introduced to convert the cell fate to that particular cell type, indicating that the ATI technique has great potential in the field of regenerative medicine.

Apart from determining the strongest DNA binders, this thesis also includes studies of nuclear proteins and DNA interactions in the context of nucleosomes, which is more similar with the real biological system. Through this new method, we are able to identify TFs that have significantly high binding specificity and affinity with the nucleosomal DNA. Some of these TFs have also been detected in the normal ATI assay, indicating that they play important roles in both transition and maintenance of the

particular cell fate (such as POU family TF in ES cells); other TFs can only be detected in the modified ATI assay using nucleosomal DNA, implying their specific roles in interacting with nucleosomes in the genome during the transition of the cell fate.

In order to measure the DNA-binding activity of cellular proteins with higher accuracy, the ATI assay should be optimized to reduce the background. For example, capillary electrophoresis could possibly be applied to study the protein-DNA interactions instead of the gel-based electrophoresis system. Besides, this massively parallel protein activity assay could also be applied to study the TF activity in single cells. This is quite promising as it can provide us with crucial information to understand the mechanism of cell fate determination in the most desired fields such as the early embryonic development of animals and the early stage of cancer development.

In conclusion, this thesis has introduced a novel technology to massively measure DNA-binding activity of all TFs in different types of cells, which is most important for us to understand the mechanism of establishment of global chromatin landscape as well as gene regulation. Based on the results, the thesis concludes that the cell identity is mainly determined by a small set of TFs with the dominant DNA-binding activity in the nucleus. The dominant TFs are responsible for establishing the kernel of the hierarchical transcriptional regulatory network and interacting with other less active TFs and cofactors to execute specific transcriptome profiles for particular cell identity.

7 ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who helped me and supported me during my PhD studies.

First and foremost, I want to express my greatest gratitude and deepest respect to my supervisor, **Prof. Jussi Taipale**. Thank you so much for giving me this precious opportunity to pursue my PhD studies in your lab! I was so lucky to find you as my PhD supervisor and get a lot of scientific training under your supervision. I also appreciate so much effort you have put on me, guiding me to the right direction when I was making mistakes. You have spent plenty of time on me, teaching me how to present our scientific work to other scientists, how to prepare scientific articles, and how to respond to the reviewers' comments. Besides, you are always supportive for anything related to my scientific training; you always encourage me to attend many international conferences, present my work to other fellows and communicate with other scientists; you are also quite supportive for my postdoctoral applications, providing quite strong recommendation and very useful suggestions. There is no doubt that nothing could be achieved without your guidance!

I 'd like to also appreciate the supervision and support from my main supervisor, **Dr. Minna Taipale**. Thank you so much for your help and support during my studies in Sweden, especially your critical comments on my manuscripts and my thesis! You are such a super nice supervisor to give me much freedom to arrange my PhD studies, taking any courses I wanted. I really appreciate so much trivial paper work from you for my studies and living in Sweden, my PhD registration, the accommodation, my residence permit application…. I cannot imagine how terrible these five years would be without your help and support! You are also quite supportive when I asked you for recommendation to pursue my post-doc positions. Finally, I really would love to, on behalf the other lab members, thank you for all the logistics work you have done for the Taipale lab.

Emma Inns, I really appreciate your help for a huge amount of administrative work you have done for me, and also your comments on my English writing! You helped me a lot on the travel expenses claim, hotel booking, logistics work for the group meetings, letter preparation, ordering the champagne and the cakes. I am aware that you have worked for so many extra hours to deal with my problems, and I appreciate it a lot!

I would like to express my sincere appreciation to the other coauthors of my ATI paper. **Dr. Arttu Jolma**, thank you so much for your suggestions to help me initiate this fantastic project, and also your useful comments on the manuscript! You are such a great scientist that I learned a lot about creativity from you. I am looking forward to meeting you and communicating with you again in the near future, and I wish you all the best with your academic career. **Dr. Inderpreet Kaur Sur**, it is a great pleasure to work with you, and I really enjoyed a lot when we were discussing science on the bus or in the lab. Thank you for all the mouse work you have done for my project, and your

precious comments on the manuscript and my thesis. I would like to thank you for your strong recommendation as well. I wish you a big future of your research group, and all the best with your family. I enjoyed playing table tennis with you little son! **Dr. Biswajyoti Sahu**, you are such a great scientist to work with, and I appreciate all the contributions and comments from you on this project. Drs. **Fan Zhong**, **Fangjie Zhu** and **Teemu Kivioja**, thank you so much for all the data analysis work you have done for my project! You guys are such great scientists, and it is a great pleasure to know you and communicate with you. I wish that we would have more communication and collaboration in the future! **Dr. Lukas M. Orre** and **Prof. Janne Lehtiö**, thank you very much for the mass spectrometry work and your great comments on this project!

There are several more colleagues and also budies I want to express my thanks to: Drs. **Yimeng Yin**, **Fangjie Zhu** and **Jian Yan**. You guys are such brilliant scientists and I learned a lot from you. Yimeng and Fangjie, I will miss the Chinese restaurant where we went for lunch, discussing about science and career every week; I will miss the yard in Huddinge where we were playing basketball; I will miss the delicious food in the best buffet restaurant in Stockholm, you guys were so generous to treat me and my girlfriend. Jian, it is great to know you and to work with you, as you are so enthusiastic and excellent. Thank you very much for all your help and suggestions on both my studies and my life in Sweden. I wish all you guys have a bright future and I am looking forward to meeting you soon!

My gratitude will be dedicated to my colleagues and previous colleagues in the Taipale group, Drs. **Otto Kauko**, **Sandeep Botla**, **Ekaterina Morgunova**, **Kashyap Dave**, **Emma Haapaniemi**, **Eevi Kaasinen**, **Åsa Kolterud**, **Bernhard Schmierer**, **Jenna Persson**, **Martin Enge**, **Jianping Liu**, **Ning Wang**, **Kazuhiro Nitta**, **Anders Eriksson** and another PhD student, **Jilin Zhang**, we have had such fantastic time in this great group. Thank you so much for all the discussions and comments! Meanwhile, I would also love to express my gratitude to the lab managers who have been working in the Taipale lab during my PhD studies. They are **Margareta Kling**, **Dr. Lijuan Hu**, **Sandra Augsten** and **Anna Zetterlund**. Thank you so much for your technical support for my project! Without your help, I won't even be able to graduate on time.

Next, I want to express my sincere gratitude to the director of the doctoral education in the department of Medical Biochemistry and Biophysics, **Prof. Elias Arnér** and the administrator **Alessandra Nanni,** and the previous director in the department of biosciences and nutrition, **Prof. Lennart Nilsson** and the administrator **Monica Ahlberg**. Thank you very much for your strong support for my graduate studies and my dissertation! Besides, I want to express my sincere appreciation to the examination board members for my halftime review, **Prof. Ernest Arenas**, **Prof. Sten Linnarsson** and **Prof. Rickard Sandberg**. Thank you so much for your great comments on my project!

I also feel most grateful that I have made so many good friends in Sweden. In the badminton group, **Zijian Qi**, **Lianhe Chu**, **Hongyi Liu**, **Wenliang Zhang**, **Peikun Sun**, **Xiaofei Li**, **Ting Jia**, **Wei Xiao**, thank you so much for making me stronger and

happier! You are my best friends forever, and I will miss every moment I have spent with you. Zijian, you are such a supportive and kind-hearted friend. I still remember the first time when we met at Frescatihallen, you were so warm-hearted while I was so cold, but it could not stop us becoming intimate friends. Apart from the badminton group, I also appreciate great help and company from other good friends. **Tiansheng Shi**, **Qing Shen**, **Meng Xie**, **Yilin Liu**, **Yang Xuan**, **Yiqiao Wang**, **Xiaopeng Hu**, **Zhe Xia**, **Zheng Chang**, **Ci Song**, **Yiting Jiang**, **Yan Chen**, **Yu Wang**, **Ying Shang**, **Xi Jiri** (**Huaqian Zhu**), **Hengsha Li**, **Erwei Zeng**, **Jiatong Li**, **Ruyue Zhang**, **Xiaolin Zhao**, **Yufei Zhu**, **Tianxiao Huang**, **Chengcheng Zhang**, **Hongying Du**, **Shuai Tan**, it is great to know all of you, and I really enjoyed the great moments with you. I hope that at some point in the future, we will gather together, playing games together. Moreover, I also want to thank other friends who are not listed here due to limited space.

Finally, I would like to express my sincere gratitude to my family. My mum and my dad, you are the best parents. Thank you so much for encouraging and supporting me, then I could have the chance to focus on my studies. I would also express my deepest apologies that I spent too little time with you. My brother Yansong, thank you for accompanying mum and dad! I am really proud of you and wish you a big future. Yashu, my sweet heart, Thank you for your understanding and your support! I was so luck to meet you, to know you and to be with you forever. I love you, my family!

Yours sincerely,
Bei Wei (韦备)

Solna, Sweden
November, 2018

8 REFERENCES

1 Schneuwly, S., Klemenz, R. & Gehring, W. J. Redesigning the body plan of Drosophila by ectopic expression of the homoeotic gene Antennapedia. *Nature* **325**, 816-818, doi:10.1038/325816a0 (1987).

2 Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987-1000 (1987).

3 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676, doi:10.1016/j.cell.2006.07.024 (2006).

4 Stadtfeld, M. & Hochedlinger, K. Induced pluripotency: history, mechanisms, and applications. *Gene Dev* **24**, 2239-2263, doi:10.1101/gad.1963910 (2010).

5 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263, doi:10.1038/nrg2538 (2009).

6 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, doi:10.1126/science.1260419 (2015).

7 Gruber, T. M. & Gross, C. A. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* **57**, 441-466, doi:10.1146/annurev.micro.57.030502.090913 (2003).

8 Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**, 105-178, doi:10.1080/10409230600648736 (2006).

9 Record, M. T., Jr., Lohman, M. L. & De Haseth, P. Ion effects on ligand-nucleic acid interactions. *J Mol Biol* **107**, 145-158 (1976).

10 von Hippel, P. H. & Berg, O. G. On the specificity of DNA-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 1608-1612 (1986).

11 Afek, A., Schipper, J. L., Horton, J., Gordan, R. & Lukatsky, D. B. Protein-DNA binding in the absence of specific base-pair recognition. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 17140-17145, doi:10.1073/pnas.1410569111 (2014).

12 Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).

13 Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* **4**, doi:10.7554/eLife.04837 (2015).

14 Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-+, doi:10.1038/nature15518 (2015).

15 Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920, doi:10.1126/science.1151526 (2007).

16 Feng, B. *et al.* Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* **11**, 197-U193, doi:10.1038/ncb1827 (2009).

17 Yamanaka, S. & Blau, H. M. Nuclear reprogramming to a pluripotent state by three approaches. *Nature* **465**, 704-712, doi:10.1038/nature09229 (2010).
18 Hata, A. & Chen, Y. G. TGF-beta Signaling from Receptors to Smads. *Cold Spring Harb Perspect Biol* **8**, doi:10.1101/cshperspect.a022061 (2016).
19 Papavassiliou, A. G., Treier, M. & Bohmann, D. Intramolecular signal transduction in c-Jun. *The EMBO journal* **14**, 2014-2019 (1995).
20 Derijard, B. *et al.* Jnk1 - a Protein-Kinase Stimulated by Uv-Light and Ha-Ras That Binds and Phosphorylates the C-Jun Activation Domain. *Cell* **76**, 1025-1037, doi:Doi 10.1016/0092-8674(94)90380-8 (1994).
21 Glickman, M. H. & Ciechanover, A. The ubiquitin-proteasome proteolytic pathway: Destruction for the sake of construction. *Physiol Rev* **82**, 373-428, doi:10.1152/physrev.00027.2001 (2002).
22 Kaiser, P., Flick, K., Wittenberg, C. & Reed, S. I. Regulation of transcription by ubiquitination without proteolysis: Cdc34/SCFMet30-mediated inactivation of the transcription factor Met4. *Cell* **102**, 303-314, doi:Doi 10.1016/S0092-8674(00)00036-2 (2000).
23 Archer, C. T. *et al.* Physical and functional interactions of monoubiquitylated transactivators with the proteasome. *J Biol Chem* **283**, 21789-21798, doi:10.1074/jbc.M803075200 (2008).
24 Melchior, F. SUMO - Nonclassical ubiquitin. *Annu Rev Cell Dev Bi* **16**, 591-+, doi:DOI 10.1146/annurev.cellbio.16.1.591 (2000).
25 Muller, S., Hoege, C., Pyrowolakis, G. & Jentsch, S. Sumo, ubiquitin's mysterious cousin. *Nat Rev Mol Cell Bio* **2**, 202-210, doi:Doi 10.1038/35056591 (2001).
26 Ross, S., Best, J. L., Zon, L. I. & Gill, G. SUMO-1 modification represses Sp3 transcriptional activation and modulates its subnuclear localization. *Mol Cell* **10**, 831-842, doi:Doi 10.1016/S1097-2765(02)00682-2 (2002).
27 Bies, J., Markus, J. & Wolff, L. Covalent attachment of the SUMO-1 protein to the negative regulatory domain of the c-Myb transcription factor modifies its stability and transactivation capacity. *J Biol Chem* **277**, 8999-9009, doi:10.1074/jbc.M110453200 (2002).
28 Kim, J., Cantwell, C. A., Johnson, P. F., Pfarr, C. M. & Williams, S. C. Transcriptional activity of CCAAT/enhancer-binding proteins is controlled by a conserved inhibitory domain that is a target for sumoylation. *J Biol Chem* **277**, 38037-38044, doi:10.1074/jbc.M207235200 (2002).
29 Desterro, J. M., Rodriguez, M. S. & Hay, R. T. SUMO-1 modification of IkappaBalpha inhibits NF-kappaB activation. *Mol Cell* **2**, 233-239 (1998).
30 Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat Rev Mol Cell Bio* **15**, 536-550, doi:10.1038/nrm3841 (2014).
31 Matsuzaki, H. *et al.* Acetylation of Foxo1 alters its DNA-binding ability and sensitivity to phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11278-11283, doi:10.1073/pnas.0502738102 (2005).

32 van der Heide, L. P. & Smidt, M. P. Regulation of FoxO activity by CBP/p300-mediated acetylation. *Trends Biochem Sci* **30**, 81-86, doi:10.1016/j.tibs.2004.12.002 (2005).
33 Lu, Q., Hutchins, A. E., Doyle, C. M., Lundblad, J. R. & Kwok, R. P. S. Acetylation of cAMP-responsive element-binding protein (CREB) by CREB-binding protein enhances CREB-dependent transcription. *J Biol Chem* **278**, 15727-15734, doi:10.1074/jbc.M300546200 (2003).
34 Boyes, J., Byfield, P., Nakatani, Y. & Ogryzko, V. Regulation of activity of the transcription factor GATA-1 by acetylation. *Nature* **396**, 594-598, doi:10.1038/25166 (1998).
35 Carr, S. M., Roworth, A. P., Chan, C. & La Thangue, N. B. Post-translational control of transcription factors: methylation ranks highly. *Febs J* **282**, 4450-4465, doi:10.1111/febs.13524 (2015).
36 Filtz, T. M., Vogel, W. K. & Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol Sci* **35**, 76-85, doi:10.1016/j.tips.2013.11.005 (2014).
37 Ozcan, S., Andrali, S. S. & Cantrell, J. E. L. Modulation of transcription factor function by O-GlcNAc modification. *Bba-Gene Regul Mech* **1799**, 353-364, doi:10.1016/j.bbagrm.2010.02.005 (2010).
38 Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669-680, doi:10.1038/nrg2641 (2009).
39 Mardis, E. R. ChIP-seq: welcome to the new frontier. *Nature Methods* **4**, 613-614, doi:DOI 10.1038/nmeth0807-613 (2007).
40 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**, W369-W373, doi:10.1093/nar/gkl198 (2006).
41 Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* **32**, W199-W203, doi:10.1093/nar/gkh465 (2004).
42 Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology* **20**, 835-839, doi:10.1038/nbt717 (2002).
43 Romer, K. A., Kayombya, G. R. & Fraenkel, E. WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids Research* **35**, W217-W220, doi:10.1093/nar/gkm376 (2007).
44 Rhee, H. S. & Pugh, B. F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol* **Chapter 21**, Unit 21 24, doi:10.1002/0471142727.mb2124s100 (2012).
45 Rhee, H. S. & Pugh, B. F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **147**, 1408-1419, doi:10.1016/j.cell.2011.11.013 (2011).
46 He, Q. Y., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology* **33**, 395-U108, doi:10.1038/nbt.3121 (2015).

47 Hampshire, A. J., Rusling, D. A., Broughton-Head, V. J. & Fox, K. R. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. *Methods* **42**, 128-140, doi:10.1016/j.ymeth.2007.01.002 (2007).
48 Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**, 3157-3170 (1978).
49 Jones, O. W. & Berg, P. Studies on the binding of RNA polymerase to polynucleotides. *J Mol Biol* **22**, 199-209 (1966).
50 Riggs, A. D., Bourgeois, S., Newby, R. F. & Cohn, M. DNA binding of the lac repressor. *J Mol Biol* **34**, 365-368 (1968).
51 Towbin, H., Staehelin, T. & Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 4350-4354 (1979).
52 Fried, M. G. & Liu, G. Molecular Sequestration Stabilizes Cap-DNA Complexes during Polyacrylamide-Gel Electrophoresis. *Nucleic Acids Research* **22**, 5054-5059, doi:DOI 10.1093/nar/22.23.5054 (1994).
53 Garner, M. M. & Revzin, A. A Gel-Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions - Application to Components of the Escherichia-Coli Lactose Operon Regulatory System. *Nucleic Acids Research* **9**, 3047-3060, doi:DOI 10.1093/nar/9.13.3047 (1981).
54 Fried, M. & Crothers, D. M. Equilibria and Kinetics of Lac Repressor-Operator Interactions by Polyacrylamide-Gel Electrophoresis. *Nucleic Acids Research* **9**, 6505-6525, doi:DOI 10.1093/nar/9.23.6505 (1981).
55 Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-1861, doi:10.1038/nprot.2007.249 (2007).
56 Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* **4**, 393-411, doi:10.1038/nprot.2008.195 (2009).
57 Berger, M. F. & Bulyk, M. L. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* **338**, 245-260, doi:10.1385/1-59745-097-9:245 (2006).
58 Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429-1435, doi:10.1038/nbt1246 (2006).
59 Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics* **36**, 1331-1339, doi:10.1038/ng1473 (2004).
60 Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* **36**, 1331-1339, doi:10.1038/ng1473 (2004).
61 Gong, W. *et al.* The development of protein microarrays and their applications in DNA-protein and protein-protein interaction analyses of

Arabidopsis transcription factors. *Mol Plant* **1**, 27-41, doi:10.1093/mp/ssm009 (2008).
62 Ho, S. W., Jona, G., Chen, C. T. L., Johnston, M. & Snyder, M. Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 9940-9945, doi:10.1073/pnas.0509185103 (2006).
63 Hu, S. H. *et al.* Profiling the Human Protein-DNA Interactome Reveals ERK2 as a Transcriptional Repressor of Interferon Signaling. *Cell* **139**, 610-622, doi:10.1016/j.cell.2009.08.037 (2009).
64 Deplancke, B., Dupuy, D., Vidal, M. & Walhout, A. J. M. A gateway-compatible yeast one-hybrid system. *Genome Research* **14**, 2093-2101, doi:10.1101/gr.2445504 (2004).
65 Reece-Hoyes, J. S. *et al.* Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nature Methods* **8**, 1059-+, doi:10.1038/Nmeth.1748 (2011).
66 Bruckner, A., Polge, C., Lentze, N., Auerbach, D. & Schlattner, U. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int J Mol Sci* **10**, 2763-2788, doi:10.3390/ijms10062763 (2009).
67 Fields, S. & Song, O. K. A Novel Genetic System to Detect Protein Protein Interactions. *Nature* **340**, 245-246, doi:DOI 10.1038/340245a0 (1989).
68 Noyes, M. B. *et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277-1289, doi:10.1016/j.cell.2008.05.023 (2008).
69 Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).
70 Cui, Y. H., Wang, Q., Stormo, G. D. & Calvo, J. M. A Consensus Sequence for Binding of Lrp to DNA. *J Bacteriol* **177**, 4872-4880, doi:DOI 10.1128/jb.177.17.4872-4880.1995 (1995).
71 Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
72 Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861-873, doi:10.1101/gr.100552.109 (2010).
73 Yin, Y. M. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, doi:ARTN eaaj223910.1126/science.aaj2239 (2017).
74 Albert, R., Jeong, H. & Barabasi, A. L. Internet - Diameter of the World-Wide Web. *Nature* **401**, 130-131 (1999).
75 Lawrence, S. & Giles, C. L. Accessibility of information on the web. *Nature* **400**, 107-109, doi:Doi 10.1038/21987 (1999).
76 Mitra, C., Kurths, J. & Donner, R. V. Rewiring hierarchical scale-free networks: Influence on synchronizability and topology. *Epl-Europhys Lett* **119**, doi:Artn 3000210.1209/0295-5075/119/30002 (2017).
77 Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512, doi:DOI 10.1126/science.286.5439.509 (1999).

78 Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98-101, doi:10.1038/nature06830 (2008).
79 Garber, M. *et al.* A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals. *Mol Cell* **47**, 810-822, doi:10.1016/j.molcel.2012.07.030 (2012).
80 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
81 Tsunaka, Y., Kajimura, N., Tate, S. & Morikawa, K. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Research* **33**, 3424-3434, doi:10.1093/nar/gki663 (2005).
82 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. *Nature* **389**, 251-260 (1997).
83 Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J Cell Biol* **83**, 403-427 (1979).
84 Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nat Rev Mol Cell Bio* **4**, 809-814, doi:10.1038/nrm1225 (2003).
85 Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**, 161-172, doi:10.1038/nrg2522 (2009).
86 Lorch, Y., Lapointe, J. W. & Kornberg, R. D. Nucleosomes Inhibit the Initiation of Transcription but Allow Chain Elongation with the Displacement of Histones. *Cell* **49**, 203-210, doi:Doi 10.1016/0092-8674(87)90561-7 (1987).
87 Knezetic, J. A. & Luse, D. S. The Presence of Nucleosomes on a DNA-Template Prevents Initiation by Rna Polymerase-Ii Invitro. *Cell* **45**, 95-104, doi:Doi 10.1016/0092-8674(86)90541-6 (1986).
88 Kayne, P. S. *et al.* Extremely Conserved Histone H-4 N Terminus Is Dispensable for Growth but Essential for Repressing the Silent Mating Loci in Yeast. *Cell* **55**, 27-39, doi:Doi 10.1016/0092-8674(88)90006-2 (1988).
89 Han, M. & Grunstein, M. Nucleosome Loss Activates Yeast Downstream Promoters Invivo. *Cell* **55**, 1137-1145, doi:Doi 10.1016/0092-8674(88)90258-9 (1988).
90 Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics* **36**, 900-905, doi:10.1038/ng1400 (2004).
91 Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626-630, doi:10.1126/science.1112178 (2005).
92 Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-366, doi:10.1038/nature07667 (2009).
93 Ozsolak, F., Song, J. S., Liu, X. S. & Fisher, D. E. High-throughput mapping of the chromatin structure of human promoters. *Nature biotechnology* **25**, 244-248, doi:10.1038/nbt1279 (2007).

94 Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898, doi:10.1016/j.cell.2008.02.022 (2008).
95 Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in S-cerevisiae. *Science* **309**, 626-630, doi:10.1126/science.1112178 (2005).
96 Polach, K. J. & Widom, J. Mechanism of Protein Access to Specific DNA-Sequences in Chromatin - a Dynamic Equilibrium-Model for Gene-Regulation. *Journal of Molecular Biology* **254**, 130-149, doi:DOI 10.1006/jmbi.1995.0606 (1995).
97 Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *Journal of Molecular Biology* **258**, 800-812, doi:DOI 10.1006/jmbi.1996.0288 (1996).
98 Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature*, doi:10.1038/s41586-018-0549-5 (2018).
99 Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions (vol 480, pg 490, 2011). *Nature* **484**, 550-550, doi:10.1038/nature11086 (2012).
100 Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics* **45**, 1198-U1340, doi:10.1038/ng.2746 (2013).
101 Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321-326, doi:10.1038/nature14192 (2015).
102 Jeong, S. *et al.* Selective Anchoring of DNA Methyltransferases 3A and 3B to Nucleosomes Containing Methylated DNA. *Molecular and Cellular Biology* **29**, 5366-5376, doi:10.1128/Mcb.00484-09 (2009).
103 MacAlpine, D. M. & Almouzni, G. Chromatin and DNA Replication. *Csh Perspect Biol* **5**, doi:ARTN a010207 10.1101/cshperspect.a010207 (2013).
104 Peterson, C. L. & Almouzni, G. Nucleosome Dynamics as Modular Systems that Integrate DNA Damage and Repair. *Csh Perspect Biol* **5**, doi:ARTN a01265810.1101/cshperspect.a012658 (2013).
105 Radman-Livaja, M. & Rando, O. J. Nucleosome positioning: How is it established, and why does it matter? *Dev Biol* **339**, 258-266, doi:10.1016/j.ydbio.2009.06.012 (2010).
106 Chen, X., Hartman, A. & Guo, S. Choosing Cell Fate Through a Dynamic Cell Cycle. *Curr Stem Cell Rep* **1**, 129-138, doi:10.1007/s40778-015-0018-0 (2015).
107 Lutter, L. C. Precise Location of Dnase-I Cutting Sites in the Nucleosome Core Determined by High-Resolution Gel-Electrophoresis. *Nucleic Acids Research* **6**, 41-56, doi:DOI 10.1093/nar/6.1.41 (1979).
108 Wu, C. The 5' Ends of Drosophila Heat-Shock Genes in Chromatin Are Hypersensitive to Dnase-I. *Nature* **286**, 854-860, doi:DOI 10.1038/286854a0 (1980).
109 Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R. & Elgin, S. C. R. Chromatin Structure of Specific Genes .1. Evidence for Higher-Order Domains of Defined DNA-Sequence. *Cell* **16**, 797-806, doi:Doi 10.1016/0092-8674(79)90095-3 (1979).

110 Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322, doi:10.1016/j.cell.2007.12.014 (2008).
111 Stalder, J. *et al.* Tissue-Specific DNA Cleavages in the Globin Chromatin Domain Introduced by Dnaase-I. *Cell* **20**, 451-460, doi:Doi 10.1016/0092-8674(80)90631-5 (1980).
112 Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* **6**, 283-289, doi:10.1038/Nmeth.1313 (2009).
113 Gilchrist, D. A. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143**, 540-551, doi:10.1016/j.cell.2010.10.004 (2010).
114 Floer, M. *et al.* A RSC/Nucleosome Complex Determines Chromatin Architecture and Facilitates Activator Binding. *Cell* **141**, 407-418, doi:10.1016/j.cell.2010.03.048 (2010).
115 Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**, 877-885, doi:10.1101/gr.5533506 (2007).
116 Horz, W. & Altenburger, W. Sequence Specific Cleavage of DNA by Micrococcal Nuclease. *Nucleic Acids Research* **9**, 2643-2658, doi:DOI 10.1093/nar/9.12.2643 (1981).
117 Dingwall, C., Lomonossoff, G. P. & Laskey, R. A. High Sequence Specificity of Micrococcal Nuclease. *Nucleic Acids Research* **9**, 2659-2673, doi:DOI 10.1093/nar/9.12.2659 (1981).
118 Lomonossoff, G. P., Butler, P. J. G. & Klug, A. Sequence-Dependent Variation in the Conformation of DNA. *Journal of Molecular Biology* **149**, 745-760, doi:Doi 10.1016/0022-2836(81)90356-9 (1981).
119 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213-+, doi:10.1038/Nmeth.2688 (2013).
120 Buenostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-U264, doi:10.1038/nature14590 (2015).
121 Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenet Chromatin* **7**, doi:Artn 33 10.1186/1756-8935-7-33 (2014).
122 Thastrom, A. *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *Journal of Molecular Biology* **288**, 213-229, doi:DOI 10.1006/jmbi.1999.2686 (1999).
123 Dechering, K. J., Cuelenaere, K., Konings, R. N. H. & Leunissen, J. A. M. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Research* **26**, 4056-4062, doi:DOI 10.1093/nar/26.17.4056 (1998).
124 Anderson, J. D. & Widom, J. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Molecular and*

Cellular Biology **21**, 3830-3839, doi:Doi 10.1128/Mcb.21.11.3830-3839.2001 (2001).
125 Kunkel, G. R. & Martinson, H. G. Nucleosomes Will Not Form on Double-Stranded-Rna or over Poly(Da).Poly(Dt) Tracts in Recombinant DNA. *Nucleic Acids Research* **9**, 6869-6888, doi:DOI 10.1093/nar/9.24.6869 (1981).
126 Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struc Biol* **19**, 65-71, doi:10.1016/j.sbi.2009.01.004 (2009).
127 Iyer, V. & Struhl, K. Poly(Da-Dt), a Ubiquitous Promoter Element That Stimulates Transcription Via Its Intrinsic DNA-Structure. *Embo Journal* **14**, 2570-2579 (1995).
128 Mavrich, T. N. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453**, 358-U327, doi:10.1038/nature06929 (2008).
129 Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772-778, doi:10.1038/nature04979 (2006).
130 Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572-576, doi:10.1038/nature05632 (2007).
131 Miele, V., Vaillant, C., d'Aubenton-Carafa, Y., Thermes, C. & Grange, T. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res* **36**, 3746-3756, doi:10.1093/nar/gkn262 (2008).
132 Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* **4**, e1000216, doi:10.1371/journal.pcbi.1000216 (2008).
133 Jin, C. Y. *et al.* Regulation of histone acetylation and nucleosome assembly by transcription factor JDP2. *Nat Struct Mol Biol* **13**, 331-338, doi:10.1038/nsmb1063 (2006).
134 Arnold, P. *et al.* Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Research* **23**, 60-73, doi:10.1101/gr.142661.112 (2013).
135 Burdon, R. H. & Adams, R. L. P. Eukaryotic DNA Methylation. *Trends Biochem Sci* **5**, 294-297, doi:Doi 10.1016/0968-0004(80)90163-2 (1980).
136 Richa, R. & Sinha, R. P. Hydroxymethylation of DNA: an epigenetic marker. *EXCLI J* **13**, 592-610 (2014).
137 Aapola, U. *et al.* Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics* **65**, 293-298, doi:10.1006/geno.2000.6168 (2000).
138 Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**, 219-220, doi:10.1038/890 (1998).
139 Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).
140 Bird, A. The essentials of DNA methylation. *Cell* **70**, 5-8 (1992).
141 Keshet, I., Lieman-Hurwitz, J. & Cedar, H. DNA methylation affects the formation of active chromatin. *Cell* **44**, 535-543 (1986).
142 Rougier, N. *et al.* Chromosome methylation patterns during mammalian preimplantation development. *Genes Dev* **12**, 2108-2113 (1998).

143 Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935, doi:10.1126/science.1170116 (2009).
144 Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929-930, doi:10.1126/science.1169786 (2009).
145 Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* **149**, 1368-1380, doi:10.1016/j.cell.2012.04.027 (2012).
146 Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397, doi:10.1038/nature10102 (2011).
147 Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 Binds to 5hmC Enriched within Active Genes and Accessible Chromatin in the Nervous System. *Cell* **151**, 1417-1430, doi:10.1016/j.cell.2012.11.022 (2012).
148 Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517-527, doi:10.1016/j.cell.2005.06.026 (2005).
149 Rao, B., Shibata, Y., Strahl, B. D. & Lieb, J. D. Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. *Molecular and Cellular Biology* **25**, 9447-9459, doi:10.1128/Mcb.25.21.9447-9459.2005 (2005).
150 Roh, T. Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Gene Dev* **19**, 542-552, doi:10.1101/gad.1272505 (2005).
151 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124, doi:10.1038/35065138 (2001).
152 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116-120, doi:10.1038/35065132 (2001).
153 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-U791, doi:10.1038/nature07107 (2008).
154 Lehnertz, B. *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* **13**, 1192-1200, doi:10.1016/S0960-9822(03)00432-9 (2003).
155 Gu, T. P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606-610, doi:10.1038/nature10443 (2011).
156 Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, doi:ARTN 241 10.1038/ncomms1240 (2011).
157 Weissmann, S. *et al.* Landscape of TET2 mutations in acute myeloid leukemia. *Leukemia* **26**, 934-942, doi:10.1038/leu.2011.326 (2012).

158 Lian, C. G. *et al.* Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma. *Cell* **150**, 1135-1146, doi:10.1016/j.cell.2012.07.033 (2012).
159 He, J. *et al.* Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. *Nature Cell Biology* **15**, 373-+, doi:10.1038/ncb2702 (2013).
160 Ang, Y. S. *et al.* Wdr5 Mediates Self-Renewal and Reprogramming via the Embryonic Stem Cell Core Transcriptional Network. *Cell* **145**, 183-197, doi:10.1016/j.cell.2011.03.003 (2011).
161 Wang, T. *et al.* The Histone Demethylases Jhdm1a/1b Enhance Somatic Cell Reprogramming in a Vitamin-C-Dependent Manner. *Cell Stem Cell* **9**, 575-587, doi:10.1016/j.stem.2011.10.005 (2011).
162 Langst, G. & Manelyte, L. Chromatin Remodelers: From Function to Dysfunction. *Genes-Basel* **6**, 299-324, doi:10.3390/genes6020299 (2015).
163 Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annu Rev Biochem* **78**, 273-304, doi:10.1146/annurev.biochem.77.062706.153223 (2009).
164 Yen, K. Y., Vinayachandran, V., Batta, K., Koerber, R. T. & Pugh, B. F. Genome-wide Nucleosome Specificity and Directionality of Chromatin Remodelers. *Cell* **149**, 1461-1473, doi:10.1016/j.cell.2012.04.036 (2012).
165 Dirscherl, S. S. & Krebs, J. E. Functional diversity of ISWI complexes. *Biochem Cell Biol* **82**, 482-489, doi:10.1139/o04-044 (2004).
166 Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. Crystal-Structure of Globular Domain of Histone H5 and Its Implications for Nucleosome Binding. *Nature* **362**, 219-223, doi:DOI 10.1038/362219a0 (1993).
167 Clark, K. L., Halay, E. D., Lai, E. S. & Burley, S. K. Co-Crystal Structure of the Hnf-3/Fork Head DNA-Recognition Motif Resembles Histone-H5. *Nature* **364**, 412-420, doi:DOI 10.1038/364412a0 (1993).
168 Cirillo, L. A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *Embo Journal* **17**, 244-254, doi:DOI 10.1093/emboj/17.1.244 (1998).
169 Chaya, D., Hayamizu, T., Bustin, M. & Zaret, K. S. Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *J Biol Chem* **276**, 44385-44389, doi:DOI 10.1074/jbc.M108214200 (2001).
170 Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell* **151**, 994-1004, doi:10.1016/j.cell.2012.09.045 (2012).
171 Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76-81, doi:10.1038/s41586-018-0549-5 (2018).
172 Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958-970, doi:10.1016/j.cell.2008.01.018 (2008).
173 Behr, R. *et al.* Mild nephrogenic diabetes insipidus caused by Foxa1 deficiency. *J Biol Chem* **279**, 41936-41941, doi:10.1074/jbc.M403354200 (2004).
174 Kaestner, K. H., Katz, J., Liu, Y. F., Drucker, D. J. & Schutz, G. Inactivation of the winged helix transcription factor HNF3 alpha affects glucose

homeostasis and islet glucagon gene expression in vivo. *Gene Dev* **13**, 495-504, doi:DOI 10.1101/gad.13.4.495 (1999).
175 Weinstein, D. C. *et al.* The Winged-Helix Transcription Factor Hnf-3-Beta Is Required for Notochord Development in the Mouse Embryo. *Cell* **78**, 575-588, doi:Doi 10.1016/0092-8674(94)90523-1 (1994).
176 Ang, S. L. & Rossant, J. Hnf-3-Beta Is Essential for Node and Notochord Formation in Mouse Development. *Cell* **78**, 561-574, doi:Doi 10.1016/0092-8674(94)90522-3 (1994).
177 Shen, W., Scearce, L. M., Brestelli, J. E., Sund, N. J. & Kaestner, K. H. Foxa3 (hepatocyte nuclear factor 3 gamma) is required for the regulation of hepatic GLUT2 expression and the maintenance of glucose homeostasis during a prolonged fast. *J Biol Chem* **276**, 42812-42817, doi:DOI 10.1074/jbc.M106344200 (2001).
178 Kuo, C. T. *et al.* Transcription factor GATA4 is required for heart tube formation and ventral morphogenesis. *Circulation* **96**, 1686-1686 (1997).
179 Morrisey, E. E. *et al.* GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. *Gene Dev* **12**, 3579-3590, doi:DOI 10.1101/gad.12.22.3579 (1998).
180 Berkes, C. A. *et al.* Pbx marks genes for activation by MyoD indicating a role for a homeodomain protein in establishing myogenic potential. *Mol Cell* **14**, 465-477, doi:Doi 10.1016/S1097-2765(04)00260-6 (2004).
181 Hsu, H. T. *et al.* Recruitment of RNA polymerase II by the pioneer transcription factor PHA-4. *Science* **348**, 1372-1376, doi:10.1126/science.aab1223 (2015).
182 Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956, doi:10.1016/j.cell.2005.08.020 (2005).
183 MacArthur, S. *et al.* Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10**, doi:ARTN R8010.1186/gb-2009-10-7-r80 (2009).
184 Levine, M. Transcriptional Enhancers in Animal Development and Evolution. *Curr Biol* **20**, R754-R763, doi:10.1016/j.cub.2010.06.070 (2010).
185 Jozwik, K. M. & Carroll, J. S. Pioneer factors in hormone-dependent cancers. *Nat Rev Cancer* **12**, 381-385, doi:10.1038/nrc3263 (2012).
186 Jain, R. K., Mehta, R. J., Nakshatri, H., Idrees, M. T. & Badve, S. S. High-level expression of forkhead-box protein A1 in metastatic prostate cancer. *Histopathology* **58**, 766-772, doi:10.1111/j.1365-2559.2011.03796.x (2011).
187 Sahu, B. *et al.* Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *Embo Journal* **30**, 3962-3976, doi:10.1038/emboj.2011.328 (2011).
188 Bass, A. J. *et al.* SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nature Genetics* **41**, 1238-U1105, doi:10.1038/ng.465 (2009).
189 Vanner, R. J. *et al.* Quiescent Sox(2+) Cells Drive Hierarchical Growth and Relapse in Sonic Hedgehog Subgroup Medulloblastoma. *Cancer Cell* **26**, 33-47, doi:10.1016/j.ccr.2014.05.005 (2014).

190 Boumahdi, S. *et al.* SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature* **511**, 246-+, doi:10.1038/nature13305 (2014).
191 Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373-1377, doi:10.1126/science.1259037 (2014).
192 Lu, Y. Z. *et al.* Functional Annotation of Risk Loci Identified Through Genome-Wide Association Studies for Prostate Cancer. *Prostate* **71**, 955-963, doi:10.1002/pros.21311 (2011).
193 Jia, L. *et al.* Functional Enhancers at the Gene-Poor 8q24 Cancer-Linked Locus. *Plos Genet* **5**, doi:ARTN e1000597 10.1371/journal.pgen.1000597 (2009).
194 Wang, Q. B. *et al.* Androgen Receptor Regulates a Distinct Transcription Program in Androgen-Independent Prostate Cancer. *Cell* **138**, 245-256, doi:10.1016/j.cell.2009.04.056 (2009).
195 Morgan, R. *et al.* Antagonism of HOX/PBX dimer formation blocks the in vivo proliferation of melanoma. *Cancer Res* **67**, 5806-5813, doi:10.1158/0008-5472.Can-06-4231 (2007).
196 Plowright, L., Harrington, K. J., Pandha, H. S. & Morgan, R. HOX transcription factors are potential therapeutic targets in non-small-cell lung cancer (targeting HOX genes in lung cancer). *Brit J Cancer* **100**, 470-475, doi:10.1038/sj.bjc.6604857 (2009).
197 Morgan, R., Plowright, L., Harrington, K. J., Michael, A. & Pandha, H. S. Targeting HOX and PBX transcription factors in ovarian cancer. *Bmc Cancer* **10**, doi:Artn 8910.1186/1471-2407-10-89 (2010).
198 Nakshatri, H. & Badve, S. FOXA1 as a therapeutic target for breast cancer. *Expert Opin Ther Tar* **11**, 507-514, doi:10.1517/14728222.11.4.507 (2007).
199 Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869-10874, doi:DOI 10.1073/pnas.191367098 (2001).
200 Badve, S. *et al.* FOXA1 expression in breast cancer - Correlation with luminal subtype A and survival. *Clin Cancer Res* **13**, 4415-4421, doi:10.1158/1078-0432.Ccr-07-0122 (2007).
201 Ademuyiwa, F. O., Thorat, M. A., Jain, R. K., Nakshatri, H. & Badve, S. Expression of Forkhead- box protein A1, a marker of luminal A type breast cancer, parallels low Oncotype DX 21-gene recurrence scores. *Modern Pathol* **23**, 270-275, doi:10.1038/modpathol.2009.172 (2010).
202 Thorat, M. A. *et al.* Forkhead box A1 expression in breast cancer is associated with luminal subtype and good prognosis. *J Clin Pathol* **61**, 327-332, doi:10.1136/jcp.2007.052431 (2008).
203 Murphy, C. L. Internal standards in differentiating embryonic stem cells in vitro. *Methods Mol Biol* **329**, 101-112, doi:10.1385/1-59745-037-5:101 (2006).
204 Morris, S. A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**, 889-902, doi:10.1016/j.cell.2014.07.021 (2014).

205 Du, Y. *et al.* Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* **14**, 394-403, doi:10.1016/j.stem.2014.01.008 (2014).
206 Huang, P. *et al.* Direct reprogramming of human fibroblasts to functional and expandable hepatocytes. *Cell Stem Cell* **14**, 370-384, doi:10.1016/j.stem.2014.01.003 (2014).
207 Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics (Oxford, England)* **25**, 3181-3182, doi:10.1093/bioinformatics/btp554 (2009).
208 Pizzi, C., Rastas, P. & Ukkonen, E. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Trans Comput Biol Bioinform* **8**, 69-79, doi:10.1109/TCBB.2009.35 (2011).
209 Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Research* **7**, 986-995, doi:DOI 10.1101/gr.7.10.986 (1997).
210 Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364, doi:10.1038/nature13992 (2014).
211 Hubner, N. C., Nguyen, L. N., Hornig, N. C. & Stunnenberg, H. G. A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *J Proteome Res* **14**, 1315-1329, doi:10.1021/pr5009515 (2015).
212 Dyer, P. N. *et al.* Reconstitution of nucleosome core particles from recombinant histones and DNA. *Method Enzymol* **375**, 23-44 (2004).
213 Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol* **8**, doi:ARTN R24 10.1186/gb-2007-8-2-r24 (2007).
214 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**, 72-U183, doi:10.1038/Nmeth.1778 (2012).
215 Kokkinopoulos, I. *et al.* Cardiomyocyte Differentiation From Mouse Embryonic Stem Cells Using a Simple and Defined Protocol. *Dev Dynam* **245**, 157-165, doi:10.1002/Dvdy.24366 (2016).
216 Zhang, J. C. *et al.* Retinoic Acid Induces Embryonic Stem Cell Differentiation by Altering Both Encoding RNA and microRNA Expression. *Plos One* **10**, doi:ARTN e013256610.1371/journal.pone.0132566 (2015).
217 Ying, Q. L., Stavridis, M., Griffiths, D., Li, M. & Smith, A. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nature biotechnology* **21**, 183-186, doi:10.1038/nbt780 (2003).
218 Gaarenstroom, T. & Hill, C. S. TGF-beta signaling to chromatin: How Smads regulate transcription during self-renewal and differentiation. *Semin Cell Dev Biol* **32**, 107-118, doi:10.1016/j.semcdb.2014.01.009 (2014).
219 Grigoryan, T. *et al.* Wnt/Rspondin/beta-catenin signals control axonal sorting and lineage progression in Schwann cell development. *Proceedings of the National Academy of Sciences of the United States of America*, doi:10.1073/pnas.1310490110 (2013).

220 Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756-766, doi:10.1038/nrg3098 (2011).
221 Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775-780 (1978).
222 Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801-813, doi:10.1016/j.cell.2013.07.034 (2013).
223 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).
224 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **6**, doi:ARTN 6186 10.1038/ncomms7186 (2015).
225 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
226 Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, doi:10.1126/science.aar3958 (2018).
227 Kato, M. *et al.* Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell* **149**, 753-767, doi:10.1016/j.cell.2012.04.017 (2012).
228 Han, T. W. *et al.* Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell* **149**, 768-779, doi:10.1016/j.cell.2012.04.016 (2012).
229 Izzo, A. & Schneider, R. Chatting histone modifications in mammals. *Brief Funct Genomics* **9**, 429-443, doi:10.1093/bfgp/elq024 (2010).
230 Li, Z. *et al.* Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608-1616, doi:10.1016/j.cell.2012.11.018 (2012).
231 Fryer, C. J. & Archer, T. K. Chromatin remodelling by the glucocorticoid receptor requires the BRG1 complex. *Nature* **393**, 88-91 (1998).
232 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**, 2227-2241, doi:10.1101/gad.176826.111 (2011).